September 2022

# Estimation of Distance of People from Video Camera without Dedicated Depth Sensors

Anonymous

# Estimation of Distance of People from Video Camera without Dedicated Depth Sensors

***Abstract*—The proposed invention utilizes the detection of human heads using the computer vision model and then camera optics information to determine the distance of each detected human head from the camera focal point. The resolution of detected distance is 0.5 meter, *i.e.*, participants are detected at discrete increments of 0.5 meters.**

***Keywords*—computer vision, head detector, depth estimation, distance estimation, video calling, video camera, machine learning**

## I. Introduction

Various applications in video conferencing systems need to detect people in the conference rooms such as 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). However, many advanced framing and tracking experiences cannot be realized without knowing how far each participant is sitting from the camera. For example, Person A is sitting at 2.5 meters from the camera and Person B is sitting at 4 meters from the camera. Knowing this information can fuel the development of various advanced framing and tracking experiences such as:

- Exclude people sitting more than X meters from the camera from framing and tracking

- If Speaker is sitting more than X meters from the camera, amplify the speech by a certain factor

- If Speaker is sitting more than X meters from the camera, exclude from the framing

- Design 2-D acoustic fence, e.g., an acoustic fence that is not active only on the horizontal FOV but also in the depth axis from the camera (that is perpendicular of camera horizontal FOV

- Create focus zones in the large conference rooms, e.g., people sitting in the zone are framed and not in the zone are not framed

- If people are sitting at a distance greater than certain meters from the camera, mute the system automatically and when any one of the participants come within the predefined distance from the camera, unmute the system.

One of the solutions of estimation of depth of people from video cameras is to add additional depth estimation / detection sensor, such as stereo vision camera, Lidar, Radar, or any other time-of-flight-based sensor that in addition to providing the visual scan of the scene also provides the depth of each pixel. This solution needs additional hardware and associated software to process the data. Also adds into the overall hardware cost of the system.

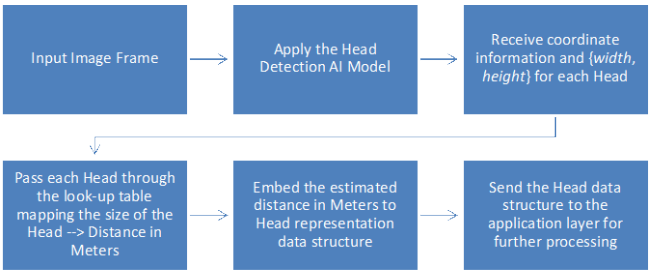
The other solution is to add monocular depth estimation models using Machine Learning technology. Two of the limitations of such approaches is high latency for processing the information frame-by-frame and the need of heavy compute if reduction in the latency is desired; and the performance of the model itself, is not as accurate as we want to be for such an application. Larger the conference rooms, courser the depth estimation they provide. One such models can be found here: https://arxiv.org/abs/1907.01341.

Further, both the approaches discussed above provide the depth of each pixel, though data rich, the information may be just overwhelming for the application where the need is only to determine how far each of the participants are sitting.

## II. Key Technology Discussions

The human head detector is an AI model that detects the location of each head in the field of view of the camera. Each head is then represented with their coordinates and height width information in the image plane. These heads are then used as an input to the distance estimation algorithm that takes {width, height} of each head and pick the best matching distance from the look-up table. The distance computation can be done in the cloud computing environment or can be implemented straight inside the cameras if has enough compute power.

The approach described here can be represented per the following computation flow:




The distance look-up table is designed using the following information.

Some metrics on the Head Size information:

**From info to right (in cm)**

| Head Breadth (exclude ears) | 1% | 5% | 50% | 95% | 99% |
|---|---|---|---|---|---|
| Men | 13.2 | 13.5 | 14.5 | 15.5 | 16 |
| Women | 12.4 | 12.7 | 13.2 | 14.2 | 15 |

| Head Height (chin to top of head) | 1% | 5% | 50% | 95% | 99% |
|---|---|---|---|---|---|
| Men | 21.2 | 21.8 | 23.2 | 24.7 | 25.5 |
| Women | 19.8 | 20.4 | 21.8 | 23.2 | 23.8 |

- So >98% of total population should fall between 12.4 cm and 16 cm breadth (with a median of 13.9 cm)
- And >98% of total population should fall between 19.8 cm and 25.5 cm head height (with a median of 22.5 cm)

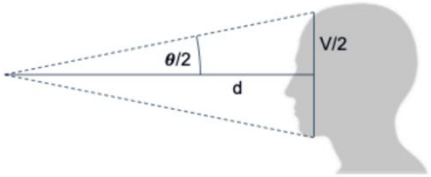
The angular extent computation is shown below.

- Angular head size, using the minimum, median and maximum head sizes, are calculated over the range of possible distance.

| | | | | Angular Extent (degrees) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Size (cm) | Distance from Camera (m) | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
| Breadth (H) | Min | 12.4 | | 17.6 | 14.1 | 11.8 | 10.1 | 8.9 | 7.9 | 7.1 | 6.5 | 5.9 | 5.5 | 5.1 | 4.7 |
| | Median | 13.9 | | 19.7 | 15.8 | 13.2 | 11.3 | 9.9 | 8.8 | 8 | 7.2 | 6.6 | 6.1 | 5.7 | 5.3 |
| | Max | 16 | | 22.6 | 18.2 | 15.2 | 13 | 11.4 | 10.2 | 9.1 | 8.3 | 7.6 | 7 | 6.5 | 6.1 |
| Height (V) | Min | 19.8 | | 27.8 | 22.4 | 18.7 | 16.1 | 14.1 | 12.6 | 11.3 | 10.3 | 9.4 | 8.7 | 8.1 | 7.6 |
| | Median | 22.5 | | 31.4 | 25.4 | 21.2 | 18.3 | 16 | 14.3 | 12.8 | 11.7 | 10.7 | 9.9 | 9.2 | 8.6 |
| | Max | 25.5 | | 35.4 | 28.6 | 24 | 20.6 | 18.1 | 16.1 | 14.5 | 13.2 | 12.1 | 11.2 | 10.4 | 9.7 |

Breadth/Width $\tan\left(\frac{\theta}{2}\right) = \frac{H/2}{d}$

Height $\tan\left(\frac{\theta}{2}\right) = \frac{V/2}{d}$




Example calculation of Angular extent at 0.4 meter distance from camera:

Min Height in CM: 19.8 → in Meter 0.198. *tan(theta/2)* = (*V*/2) / *d* = (0.198 / 2) / 0.4 → *tan(theta/2)* = 0.2475 → *theta/2* = 13.90 degrees → *theta* = 27.8 Degrees.

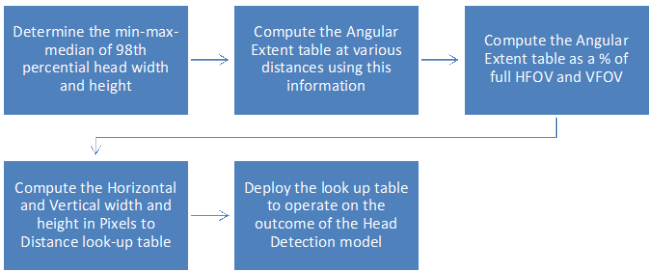
By the same logic, the entire table can be computed.

If we have a video device with $83^0$ horizontal field of view (FOV) and $53^0$ vertical FOV, we can represent these angular extents as % of full FOV.

For example, vertical $27.8^0$ are 52.45% (27.8/53 * 100) of $53^0$.

The full table of angular Extent as a % of full Horizontal and Vertical FOV is given below:

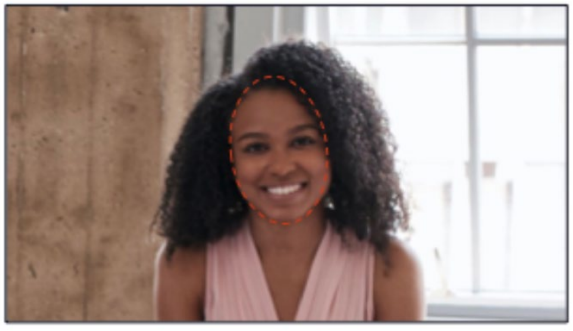
Now using the full horizontal FOV in pixels and vertical FOV in pixels, this table can be converted to head size in number of pixels. To realize the proposed invention, it is necessary to generate the look up table that converts # pixels height and width information to distance in meters. The approach for generating the look up table can be represented per the following logical process and computations flow.




- The table below shows the angular extent of the subject's head as a % of the full FOV.

| | | | | % of Frame | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Size (cm) | Distance from Camera (m) | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
| Breadth (H) | Min | 12.4 | | 21.2% | 17.0% | 14.2% | 12.2% | 10.7% | 9.5% | 8.6% | 7.8% | 7.1% | 6.6% | 6.1% | 5.7% |
| | Median | 13.9 | | 23.7% | 19.0% | 15.9% | 13.6% | 11.9% | 10.6% | 9.6% | 8.7% | 8.0% | 7.3% | 6.9% | 6.4% |
| | Max | 16 | | 27.2% | 21.9% | 18.3% | 15.7% | 13.7% | 12.3% | 11.0% | 10.0% | 9.2% | 8.4% | 7.8% | 7.3% |
| Height (V) | Min | 19.8 | | 52.5% | 42.3% | 35.3% | 30.4% | 26.6% | 23.8% | 21.3% | 19.4% | 17.7% | 16.4% | 15.3% | 14.3% |
| | Median | 22.5 | | 59.2% | 47.9% | 40.0% | 34.5% | 30.2% | 27.0% | 24.2% | 22.1% | 20.2% | 18.7% | 17.4% | 16.2% |
| | Max | 25.5 | | 66.8% | 54.0% | 45.3% | 38.9% | 34.2% | 30.4% | 27.4% | 24.9% | 22.8% | 21.1% | 19.6% | 18.3% |

Example with Head Height at 45% of frame is shown below.



Example with 45% Head Height

Example image with full vertical field of view of 2160 pixels and head height of 840 pixels is shown below. Using the look up table, it can be estimated that the person is about 0.7 meter from the camera.

- Using the full resolution of the camera (3840 horizontal pixels x 2160 vertical pixels), the extent of the head in pixels can be determined from the % of the frame
- The table below shows the size of the human head in number of pixels.

| | | | | # pixels | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Size (cm) | Distance from Camera (m) | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
| Breadth (H) | Min | 12.4 | | 814 | 653 | 545 | 468 | 411 | 365 | 330 | 300 | 273 | 253 | 234 | 219 |
| | Median | 13.9 | | 910 | 730 | 611 | 522 | 457 | 407 | 369 | 334 | 307 | 280 | 265 | 246 |
| | Max | 16 | | 1044 | 841 | 703 | 603 | 526 | 472 | 422 | 384 | 353 | 323 | 300 | 280 |
| Height (V) | Min | 19.8 | | 1134 | 914 | 762 | 657 | 575 | 514 | 460 | 419 | 382 | 354 | 330 | 309 |
| | Median | 22.5 | | 1279 | 1035 | 864 | 745 | 652 | 583 | 523 | 477 | 436 | 404 | 376 | 350 |
| | Max | 25.5 | | 1443 | 1166 | 978 | 840 | 739 | 657 | 592 | 538 | 492 | 456 | 423 | 395 |

***Note**: pixels based on 3840 x 2160 full resolution – different limits apply to downscaled streams*

## III. ADVANTAGES

1. Sensor-less. No need for any additional hardware, firmware specific to the hardware, and data manipulation software associated with the hardware to convert to distance in meters for each Head.
2. No added cost. Available for free with the existing Head detection algorithm with low compute based on the look up table.
3. Fully software enabled AI-model based Head distance inference.
4. Creates value in various applications described in the beginning of the invention.
5. Approximation of room dimensions and product recommendations based on the distances of heads.