Supplementary Material: Active Crowd Analysis for Pandemic Risk Mitigation for Blind or Visually Impaired Persons

Samridha Shrestha[1,2], Daohan Lu[1,3], Hanlin Tian[1,3], Qiming Cao[1,3], Julie Liu[1,2], John-Ross Rizzo[3], William H. Seiple[4], Maurizio Porfiri[3], and Yi Fang⋆[1,2,3]

[1] NYU Multimedia and Visual Computing Lab, Abu Dhabi and New York
[2] New York University Abu Dhabi, Abu Dhabi 129188, UAE
[3] New York University, New York, NY 10012, USA
[4] Lighthouse Guild, New York, NY 10023, USA

⋆ indicates corresponding author : yfang@nyu.edu

1 Centroid Tracking Algorithm

Algorithm 1: Bounding Box Centroid Tracking Algorithm for motion tracking

```
// bbox = bounding box, cent = centroid
// bbox_cent_map = f : bbox_id → bbox_centroid
// MAX_FRAME = max number of frames before dropping missing
   object
Input: Newly detected bbox centroids
Output: Updated bbox_cent_map
if No bbox detected then
    foreach bbox_i in bbox_map do
        bbox_i absence_count += 1
        if bbox_i absence_count > MAX_FRAME then
            deregister bbox id_i
    return Updated bbox_centroid_map
if No previously tracked centroids then
    foreach centroid_c in new bbox centroids do
        register bbox with centroid_c
else
    D_{prev_c,new_c} ← euclidean dist between each prev cent & new cent
      pair
    sort distances in D_{prev_c,new_c} in an ascending order row-wise &
      column-wise
    foreach row, col in D_{prev_c,new_c} do
        cur_bbox_id ← bbox_cent_map[row]
        bbox_cent_map_{cur_bbox_id} ← new_c[col]
        bbox_absence_count_{cur_bbox_id} ← 0
        used_{row,col} ← row, col
    unused_prev_bbox ← all_{row,col} − used_{row,col}
    if len(new_c) < len(prev_c) then
        foreach id_i in unused_prev_bbox do
            id_i absence_count += 1
            if id_i absence_count > MAX_FRAME then
                deregister bbox id_i
    else
        foreach id_i in unused_prev_bbox do
            register bbox id_i
return Updated bbox_centroid_map
```

2 Crowd Risk Evaluation Function

The individual crowd risk evaluation function defined in Section 4.5 in the main text converts the three-dimensional feature vector $\mathbf{c} \equiv s, d, v$ of the individual crowd $\mathbf{c}$ into a risk score $r(\mathbf{c})$. This crowd *riskiness* formula is mathematically defined as:

$$\begin{aligned} \mathbf{c} &\equiv s, d, v \\ r(\mathbf{c}) &= f(s, d, v) \end{aligned} \tag{1}$$

where $r : \mathbb{R}^3 \rightarrow \mathbb{R}^1 \in (0, 1)$ is a function that converts the size s (number of people), average distance d, and motion v (signed real-number velocity) of a crowd into a real number representing the risk of the crowd. We define the individual crowd risk function r according to official social distancing guidelines such as those set by the CDC [2]. These guidelines would dictate that the risk level is significantly elevated for each visible human or crowd if their distance d is less than 6 feet or 1.8 meters. The individual crowd-risk $r(\mathbf{c})$ is also directly proportional to the crowd's size and velocity towards the user and inversely proportional to the distance of the crowd from the user. The detailed individual crowd-risk function is then calculated as:

$$f(s, d, v) = \frac{Dist_c}{d} \times (Size_c \times s + min(0, d - dist_{social}) + Vel_c \times v) \tag{2}$$

where $Dist_c$, $Size_c$, and Vel_c represent the scaling or threshold factors for the crowd distance, size, and velocity respectively. Higher values of these thresholds indicate a lower tolerance for the risks associated with visible crowds. The $dist_{social}$ is the social-distance recommended by the associated health authorities which according to the CDC is 1.8 meters [2] in this case. The values for these scaling factors are experimentally determined to best suit the needs and tolerance of the BVI user and any other distancing guidelines. The $Dist_c$ scaling factor and the distance d affect all the other factors directly because we do not care if the crowd has a lot of people if the crowd is at a great distance from the BVI user. Additionally, other metrics for risk evaluation that conform to the official advisory regarding different pandemics can be also be selected for risk calculation.

The overall riskiness $R(\mathbf{c})$ is then simply defined as the sum of individual riskiness of crowds:

$$R(\mathbf{c}) = \sum_{\mathbf{c} \in \mathbf{C}} r(\mathbf{c}) \tag{3}$$

where $\mathbf{C}$ is the set of all visible crowds.

3 Bounding Box Regression Distribution

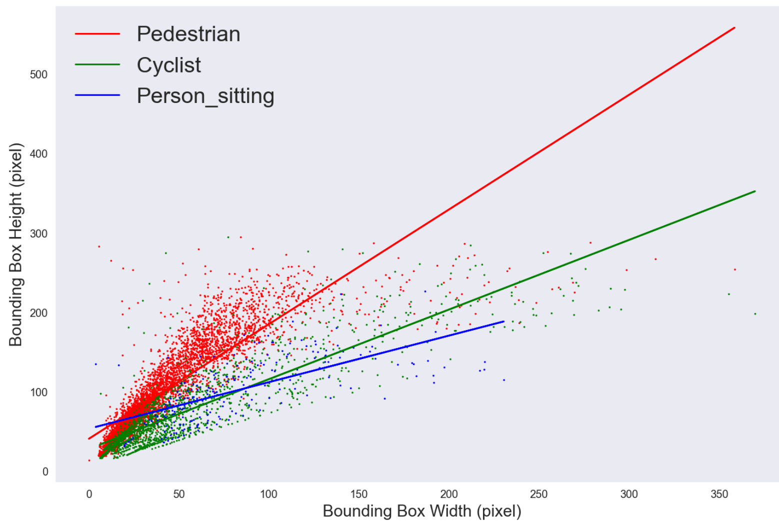

Fig. 1: Regression of bounding box height against widths to show the different distributions of the Pedestrian, Cyclist, and Person Sitting classes

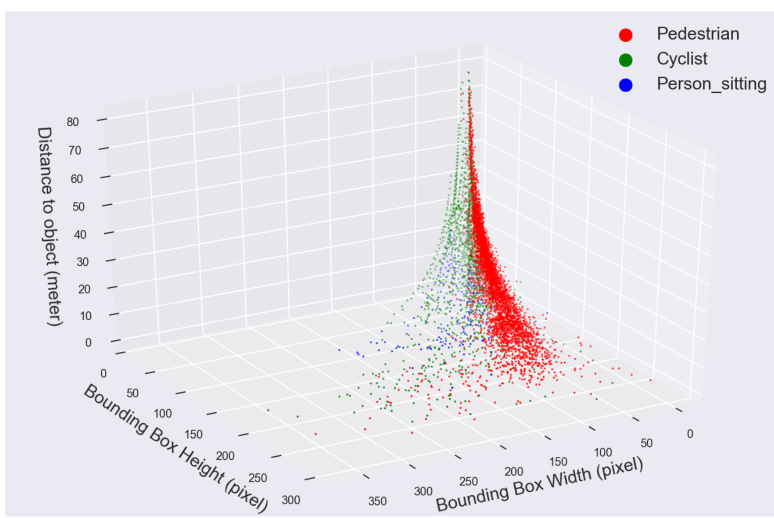

Fig. 2: Plot of bounding box height, width, and object distance to show the different distributions of the Pedestrian, Cyclist, and Person Sitting classes

4 Distance Regression Evaluation Metrics

The distance regression evaluation metrics include: the Mean Squared Error (*MSE*), the root of the mean squared error (*RMSE*), the root of the mean squared error computed from the log of the predicted distance and the ground truth distance ($RMSE_{log}$), the absolute relative difference in distances (*Abs Rel*), and the squared relative difference in distances (*Squa Rel*). If the predicted distance is d and the ground truth distance is d^{gt}, then the distance regression metrics are calculated as:

$$\text{MSE (linear)} : \frac{1}{N}\sum_{d\in N}\left\|d_i - d_i^{gt}\right\|^2 \tag{4}$$

$$\text{RMSE (linear)} : \sqrt{\frac{1}{N}\sum_{d\in N}\left\|d_i - d_i^{gt}\right\|^2} \tag{5}$$

$$\text{RMSE (log)} : \sqrt{\frac{1}{N}\sum_{d\in N}\left\|\log d_i - \log d_i^{gt}\right\|^2} \tag{6}$$

$$\text{Abs Relative Difference} : \frac{1}{N}\sum_{d\in N}\frac{|d_i - d_i^{gt}|}{d_i^{gt}} \tag{7}$$

$$\text{Squared Relative Difference} : \frac{1}{N}\sum_{d\in N}\frac{\left\|d_i - d_i^{gt}\right\|^2}{d_i^{gt}} \tag{8}$$

5 Motion Tracking Evaluation Metrics

The motion tracking evaluation metrics use the *CL*assification of *E*vents, *A*ctivities, and *R*elationships (*CLEAR*) metrics [1]. *MOT* bench-marking is a difficult task [4] due to reasons such as lack of predefined training and testing data, and numerous ambiguous evaluation metrics with free parameters. The CLEAR metrics tackles the previous bench-marking issues by introducing standardized metrics such as the Multiple Object Tracking Accuracy (*MOTA*) which is calculated as:

$$\text{MOTA} : 1 - \frac{\sum_t (FN_t + FP_t + IDsW_t)}{\sum_t GT_t} \tag{9}$$

where t is the frame index, FN_t and FP_t are the number of false negatives and false positives for the tth frame. $IDsW_t$ represents the mismatch or the identity switch error when tracking multiple objects. The *CLEAR* metric also includes the Multiple Object Tracking Precision(*MOTP*) calculated as:

$$\text{MOTP} : \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \tag{10}$$

where $d_{t,i}$ is the bounding box overlap of the target i with respect to its corresponding ground truth object and c_i number of matches for the frame t. Additional tracking metrics include the Mostly Tracked (MT), Mostly Lost(ML), False Positives (FP), False Negatives(FN), and Identity swaps ($IDsW$). We use the same definitions for MT and ML used in [4]. MT and ML measure the tracking quality for each target object's track across time frames. A target is stated to be mostly tracked if it is tracked for at least 80% of its visible path. Alternatively, a mostly lost target is labeled as such if it was only recovered for less than 20% of its total life-span. MT and ML are reported as ratios of mostly tracked and mostly lost objects to the total number of ground truth tracks. If the ID of tracked target changes during tracking, neither of the MT and ML metrics are affected. $IDsW$ on the other hand represents an tracked-object identity mismatch error (i.e. if the ID of a tracke object changes during its life-span).

6 System Specifications

We provide the system details for a fair benchmark comparison of our results in the future. The final implementation of the Active-Crowd-Analysis System however will not be contingent on the provided specifications as we will using the System-On-Chip hardware components of smartphones. Out system will use a quantized neural network with a MobileNet-V2 [5] backbone for human-detection. A meticulous breakdown of object detection performance on smartphones using MobileNet-V2 is discussed in [3] along with benchmarks for image-recognition performance for different phones and neural-network models.
The desktop system we use for training and testing our models has the following specifications.

- CPU: Intel(R) Core(TM) i99900K CPU 3.60GHz 64
- RAM: 62 Gib System Memory
- Graphics: 7979 MiB GeForce RTX 2080

Acknowledgement

The authors gratefully acknowledge the financial support from the NYUAD Institute (Research Enhancement Fund - RE132).

References

1. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008). https://doi.org/10.1155/2008/246309
2. CDC: Social distancing, quarantine, and isolation (May 2020), https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html
3. Ignatov, A., Timofte, R., Chou, W., Wang, K., Wu, M., Hartley, T., Van Gool, L.: AI Benchmark: Running Deep Neural Networks on Android Smartphones. arXiv:1810.01109 [cs] (Oct 2018), http://arxiv.org/abs/1810.01109, arXiv: 1810.01109
4. Milan, A., Leal-Taixe, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv:1603.00831 [cs] (May 2016), http://arxiv.org/abs/1603.00831, arXiv: 1603.00831
5. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381 [cs] (Mar 2019), http://arxiv.org/abs/1801.04381, arXiv: 1801.04381