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

Abstract. During pandemics like COVID-19, social distancing is essential to combat the rise of infections. However, it is challenging for the visually impaired to practice social distancing as their low vision hinders them from maintaining a safe physical distance from other humans. In this paper, we propose a smartphone-based computationally-efficient deep neural network to detect crowds and relay the associated risks to the Blind or Visually Impaired (BVI) user through directional audio alerts. The system first detects humans and estimates their distances from the smartphone's monocular camera feed. Then, the system clusters humans into crowds to generate density and distance maps from the crowd centers. Finally, the system tracks detections in previous frames creating motion maps predicting the motion of crowds to generate an appropriate audio alert. Active Crowd Analysis is designed for real-time smartphone use, utilizing the phone's native hardware to ensure the BVI can safely maintain social distancing.

Keywords: active crowd analysis, visually impaired, human detection, crowd density, crowd distance, crowd motion, crowd-risk alert, pandemic risk mitigation

1 Introduction

The World Health Organization estimates that there are about 39 million blind and 246 million visually-impaired individuals in the world [5, 66]. Numerous reports [28,36,38,63], have stated that even before the start of the current pandemic, low vision already posed significant challenges to the visually impaired individuals in conducting their day-to-day activities. Recent surveys conducted by the American Foundation for the Blind [10] and the Canadian Council for the Blind [26] found that the COVID-19 outbreak profoundly exacerbated those

[⋆] indicates corresponding author : yfang@nyu.edu existing hurdles for the BVI. These challenges included issues in public navigation, transport, and shopping all while avoiding crowds making social-distancing difficult. BVI individuals face these barriers as they have to rely on physical sensation significantly more than the normal-sighted in their everyday lives to locate objects or navigate environments. This reliance on physical touch prevents effective social distancing and puts the sightless at an elevated risk of contracting viruses by being in the vicinity of infected individuals [67]. To ameliorate the reduced perceptive range of the BVI and to mitigate the pandemic health risks, in this paper, we propose Active Crowd Analysis, a system to augment the BVI's environment perception to detect nearby visible crowds and maintain social-distancing in a more intuitive, safe, and independent manner.

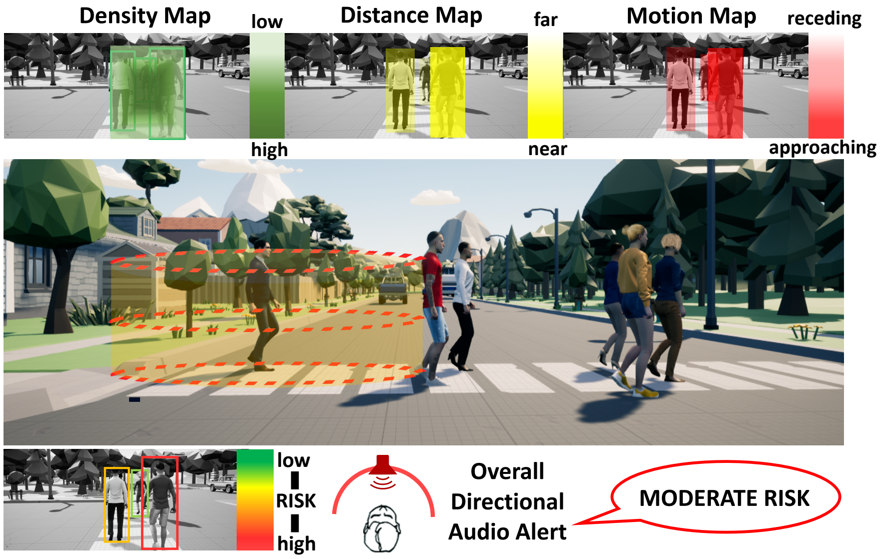


Fig. 1: Active Crowd Analysis System to detect crowds, generate density, distance, and motion maps to finally relay a risk alert to the BVI user

Active Crowd Analysis holistically integrates the density, distance, and the motion of the visible crowd in-front of a BVI person to evaluate the riskiness of crowds and to help the user avoid crowds through audio warnings (Fig. 1). The system is designed for use in a smartphone that is mounted in front of the BVI person with a lanyard. Allowing for wider access, our system requires no specialized hardware except for a standard CPU/GPU and camera-enabled smartphone along with a headphone for audio feedback. We recommend a bone-conduction headphone in specific as such headphones do not obstruct the normal hearing of a visually impaired person while still providing the necessary audio guidance to avoid crowds. The system consists of a backbone feature network to extract features from images from the smartphone camera which is passed to a human detector to detect crowds and create a crowd density map (Fig. 2). The bounding box coordinates of the detected humans are then sent to a distance regressor network to calculate the distances to detected individuals to create a distance map (Fig. 2). Using the detected human and their distances from multiple frames, the system finally generates a motion map (Fig. 2). Using information from the three previous maps, a crowd-risk module then alerts the BVI user of any risk from visible crowds nearby through the bone-conduction headphones as spatialized directional audio (Fig. 6). The system uses computationally-light neural networks that were designed for real-time smartphone use [59] to provide an active and reliable risk mitigation solution for the BVI during a pandemic.

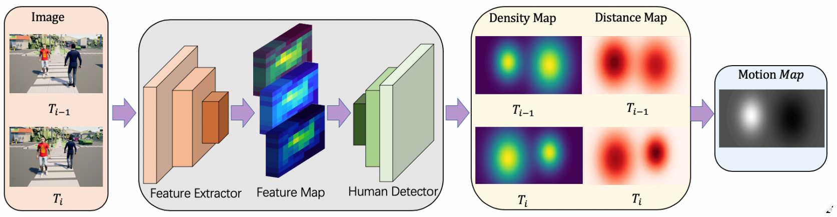


Fig. 2: Crowd-Risk Alert classifies risk-inducing crowds and sends the BVI a concise verbal alert with spatial audio.

1.1 Contributions

In short, our contributions can be summarized as follows

- Crowd Density Analysis: Calculates the density of the visible crowd in-front of the BVI individual using an efficient density-based clustering algorithm for crowd-density evaluation.
- Crowd Distance Analysis: Calculates the distance to each detected human in-front of the BVI person for crowd-distance evaluation.
- Crowd Motion Analysis: Calculates the perceived change in distance to the detected humans to detect motion for crowd-motion evaluation.
- Crowd Risk Analysis: Integrates the risk scores from the density, distance, and motion evaluations to calculate an aggregated risk that is relayed to the BVI user through spatialized 2D audio.

2 Needs of the BVI during a Pandemic

Surveys in [2, 10, 13, 26] stated that one of the paramount needs of the visually impaired during pandemics was maintaining social-distancing while approaching crowded areas where there is a greater risk of contracting the pandemic virus. During a pandemic, the BVI not only have to minimize physical touching for environment exploration but also stay away from crowds to avoid contracting contagions. Therefore, the BVI require real-time information about the surrounding environment such as the existence of crowds to actively avoid them and prevent the transmission of the disease. [11,12,33,49,52] also show that the BVI community is more receptive to lightweight and small wearable solutions with ubiquitous availability (i.e. smartphones) that enable them to perform their daily activities safely, reliably, and independently.

3 Related Work

3.1 Related Assistive Technology for the BVI

Most hardware-based assistive technology solutions for the BVI have had dire adoption rates [23, 44, 57] owning to significant drawbacks that include high cost, steep learning curves, and heavy and unwieldy hardware. The most widely adopted commercial sensory augmentation devices generally cost upwards of thousands of US dollars due to the high cost of specialized hardware [20,25,30, 32,45,65]. Other hardware-based assistive devices often cause discomfort to the user after long use due to carrying additional hardware such as batteries or cameras. Besides, such existing assistive technology solutions were not designed to address the pandemic-specific needs of the BVI. Most assistive technologies focus on limited applications like outdoor navigation [29,37,41,42,57]. Vision substitutes [4,6,15,35] on the other hand do not provide health and safety assistance to the BVI in the context of a pandemic. In contrast, software solutions that run on smartphones are more affordable and accessible for the BVI and include applications like Microsoft Seeing AI [46] and BlindSquare [3]. Unfortunately, these technologies also do not address pandemic risk reduction for the BVI like helping them to maintain social-distance from crowds. These systems also potentially fail due to the lack of onboard visual processing and the dependency on online visual computing platforms (i.e. Microsoft Seeing AI).

3.2 Related Work in Crowd Density, Distance, and Motion Analysis

Human Density Estimation and Motion Tracking: Previous work in crowd density has necessarily involved real-time human detection and motion tracking. [34] use the You Only Look Once V3 (YOLO-V3) object detection algorithm [54] to detect people and implemented background-subtraction with Gaussian Mixture Models (GMM) and contour heat-maps to analyze crowd densities. [62] used the RGB color features of an image and similar background subtraction between frames to filter background noise to detect and track moving objects in video scenes. However, these methods are only applicable to static video surveillance since background subtraction is inapplicable when the camera is moving. Therefore, these methods for people detection and density analysis would fail when a BVI user is moving around. [9] also implement a background subtraction method as a preprocessing step to detect silhouettes of people in still images using a graph-cut segmentation plan but cannot be used in real-time video sequences. [58] proposed an improved algorithm for object motion estimation in videos where they use GMMs for background subtraction and noise cancellation along with an optical flow algorithm to track objects. Background subtraction methods as mentioned previously restrict the BVI user to a stationary position. Besides, background subtraction methods for videos have deteriorated performance if the person(s) detected in the image is not moving since the foreground selection is based on the movement of the target (human) and a static background. Other methods for tracking people used R-CNN [22] and Faster R-CNN [8] object detection networks to detect people and Euclidean distance based object association between subsequent frames to track people. These methods generate accurate object detections. However, R-CNN [22] has a multi-stage region proposal selective search algorithm that generates 2000 regions to be fed into the neural network classifier while Faster R-CNN [55] still uses a region proposal network which drastically reduces the object detection speed. This means these methods are not suitable for real-time crowd detection in smartphones or embedded-devices.

Human Distance Estimation: Crowd-distance estimation is linked to depth estimation and segmentation of objects in images. Depth segmentation methods usually rely on stereo images from two cameras where the distance information is only calculable for the overlapping fields of view between the cameras [51]. For example, [18] used a scene-geometry based method to use depth cues from stereo images to track moving pedestrians from a moving platform. However, stereo images imply a dual-camera requirement which is undesirable as the BVI user would require two smartphone cameras set up infront of them or multiple-camera embedded devices. For distance estimation from monocular images, various works have been documented. [56, 64] use an object distance estimation algorithm for monocular images based on inverse-perspective-mapping (IPM) of the camera's 2D image into a bird's eye view coordinate using the camera parameters (focal length, height, etc). Despite working from a single camera, IPM has significant disadvantages as it fails to accurately predict distance for objects on the borders of the image, or curved surfaces. IPM also requires constant calibration (such as the white markings on a road) and a static height for the camera [64] which cannot always be ensured for our specific use with the BVI. [24] proposed another monocular image distance estimation method that used a Support Vector Regressor on the bounding box width and height to predict the distance to the object while in [27], the authors used DistNet to predict the distance from the bounding box features and used a CNN based object detection model (YOLO) to detect the objects themselves. [68] provided an improved and novel method to use features extracted by a neural network such as Resnet or VGG [61] to directly estimate the distance to the detected object. However, these approaches are not suitable for real-time applications on mobile systems such as smartphones due to their excessive memory and computation requirements.

4 Active Crowd Analysis

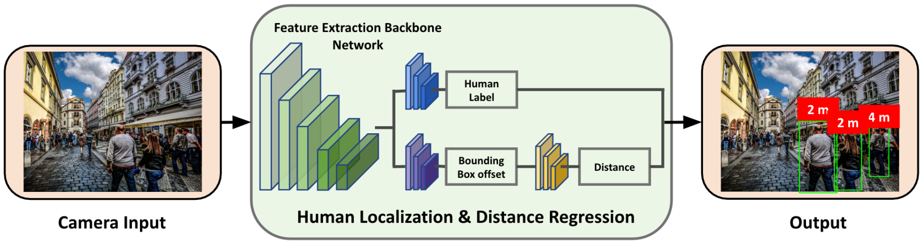

Feature Extraction Backbone Network
Human Label
Bounding Box offset
Distance
Human Localization & Distance Regression
Camera Input
Output
2 m
2 m
4 m

Fig. 3: Human detection and distance estimation network

Active Crowd Analysis is a smartphone-based assistive technology designed for the smartphone's limited memory to achieve real-time human detection, tracking, and distance estimation, traditionally reserved for GPU-based desktop computers. Active-Crowd Analysis identifies and alerts the BVI user of risk-inducing events related to crowds with two sequential modules, Crowd Analysis, and Crowd-Risk Alert. First, the Crowd Analysis module detects all visible people on the camera and clusters people into crowds to generate the crowd density map, crowd distance map, and crowd motion map, as depicted in Fig. 2. The Crowd-Risk Alert module then analyzes the information from the crowd density, distance, and motion analysis to estimate the ”riskiness” of crowds per advised health guidelines, such as social distancing [7]. The system finally relays a summary of the crowd's risk analysis (i.e. the level of risk) to the BVI user through a spatialized 2D audio.

For real-time processing on a mobile phone, we develop a fast detection algorithm as shown in Fig. 3 to compute the locations and distances of people on the smartphone camera. It is realized with 1). designing a lightweight feature extractor to simultaneously acquire foundational visual features for different visual tasks in Scene Crowd Analysis 2). designing a shared people-object detector to detect both people and objects simultaneously for Crowd detection, distance, density, and motion map generation. An overview of the people detection algorithm is shown in Fig. 2. The shared feature extractor is modeled based on MobileNet-V2 [59] and further simplified for real-time inference in smartphones. The people-object detector, which classifies the feature maps from the shared feature extractor, is based on SSDLite [60], an object detection algorithm specifically optimized for mobile devices. The detector identifies multiple objects in the image including person and non-person objects, but the Crowd Analysis filters out all the non-human detections.

4.1 Human and Crowd Detection

Detecting humans in the smartphone camera feed is an object detection task. This computer vision task is relatively easy given that many object detection algorithms such as SSD [40] and YOLO [53] can reliably pick up people and objects in an image. However, we require object detection systems that can reliably detect humans in real-time when operated from smartphones. We use a backbone feature extraction network that is based on the MobileNet-V2 network [59] and a bounding box regressor and classification network based on SSD Lite [60]. The outputs from the backbone feature extractor are of sizes 20*20*96, 10*10*1280, 5*5*512, 3*3*256, 2*2*256, and 1*1*64 with attached $4, 6, 6, 6, 4$, and 4 anchor boxes for each output feature map respectively. We also use a non-max suppression threshold of 0.5 to remove multiple detections of the same object. Our loss function is a weighted combination of losses from the object localization (loc) and classification tasks (cls) based on the multi-box detection for multiple classes used in [17, 60].

$$L(x,c,l,g) = \frac{1}{N}(L_{cls}(x,c) + \alpha L_{loc}(x,l,g)) \tag{1}$$

where N is the number of matched default box priors. The loss is set to 0 if N is also 0. x is an indicator variable that is set to 1 if the default prior box is matched to the determined ground truth box and 0 otherwise. c is the class confidence score, l and g represent the predicted and ground truth bounding box parameters (center offsets; bounding box width and height) respectively. L_{loc} is the localization loss which is the smooth L1 loss between the predicted (l) and ground-truth box (g) parameters. α is a hyper-parameter that balances the weights of the losses and is determined through cross-validation. L_{cls} is the classification Softmax loss computed over multiple classes. However, after detecting an object, we drop all classes except for the person class in our human detection pipeline.

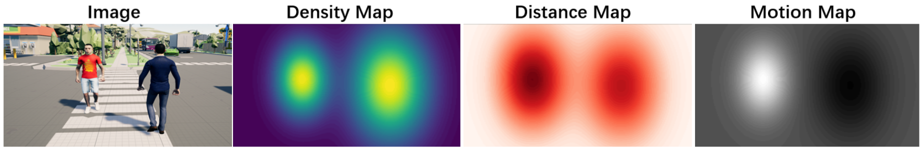


Fig. 4: Illustrations of Crowd Density Map, Crowd Distance Map and Crowd Motion Map.

4.2 Crowd Density Analysis

To create density maps, we use bounding-box parameters ($center : (x, y), width : w, height : h$) generated by the human detection module for each detection. A distance regressor network uses the bounding-box w and h to predict the distance, d to each detected human. We then use the bounding-box centers and distances, (x, y, d) to represent individual persons and form centroid clusters based on 3D euclidean distances between bounding-box centers using a recursive density-based clustering algorithm, DBSCAN [19]. DBSCAN is well suited to our clustering task compared to an algorithm like K-means clustering which is affected by noise and requires the number of clusters apriori. DBSCAN does not require the number of clusters as a hyper-parameter input, instead, it requires the epsilon (*eps*) parameter which is the maximum distance between two points in the same cluster. The most suitable *eps* was experimentally determined to be 18 after re-scaling the dimensions of the input images to $[0, 100]$ for increased generalizability. To generate the density maps, we employ a GMM model for contour heat-map similar to what was used in [34] albeit without using any background subtraction method. The number of components in the GMM model is simply the number of crowd-clusters that was previously computed with DBSCAN.

4.3 Crowd Distance Analysis

For crowd distance map, it is difficult to directly obtain the distances to detected objects because most object detection algorithms do not compute the distance of the detected objects to the camera. To acquire the distances to humans, we use the proportional geometric relationship between the height and the width of the bounding box to the actual size of the person. Such distance estimation of people on a monocular image is intuitively justified as humans also have well-defined shapes when captured by a camera, so their on-camera appearance (size and look) gives a good estimate of their distances. We use a fully-connected deep neural network to regress for the distance to detected human trained on labeled ground-truth data. As shown in Fig. 3, we predict the location and size of the bounding-box and use the width and height of the predicted bounding box to calculate the distance to the detected humans. The Distance regressor network is a deep neural network with five fully connected layers of size 2, 6, 4, 2, 1, experimentally determined to be optimal after hyper-parameter optimization. The network has LeakyRELU activation (slope = -0.01) functions between all subsequent layers except for the final layer which has a Softplus activation ensuring all final distance predictions are positive. We train our distance regression network with a supervised Mean Squared Error loss function presented below:

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

where d_i is predicted distance and d_i^{gt} is the ground-truth distance to the detected human. The network outputs a distance for every detected human in the input image and assigns the distance to the bounding box center.

4.4 Crowd Motion Analysis

After we detect humans and calculate the distances to each human and cluster center, we can start to track the motion of each detected human or cluster center. In general pedestrian settings, people move relatively slowly as viewed on a camera, so there is a considerable frame-to-frame overlap of their locations. Thus, by comparing the distances of the tracked cluster centers or humans from frame to frame, we can reliably estimate the motion of such crowds or humans as viewed from the BVI user's smartphone camera. By calculating the change in the distance values between consecutive frames, we can calculate the motion of crowds. This velocity will indicate whether detected humans or crowds are moving towards or away from the BVI user and can be used to create the motion map. Motion tracking, however, is essential for estimating motion between frames. To track the detected humans between subsequent frames, we employ a simple bounding box centroid tracking algorithm that links previously-detected bounding boxes with newly detected bounding boxes with the smallest euclidean distance. To suppress noisy detections, we only start tracking humans once they have been visible for a small number of frames and drop missing human detections once they have been absent for several frames (set to 50 in our case). The centroid tracking algorithm is discussed in more detail in Section 1 of the supplementary material. (Sup. Sec. 1)

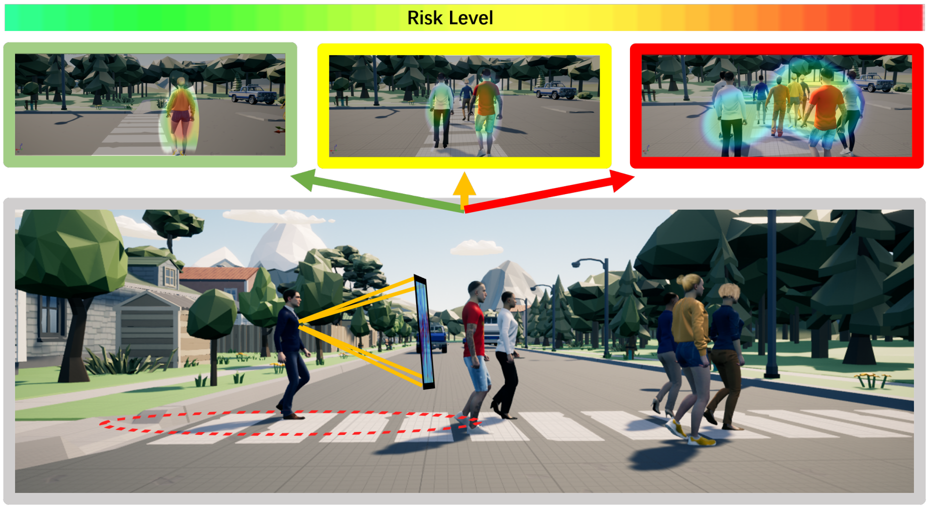


Fig. 5: Active Crowd Analysis generates a straightforward risk level for detected crowds, where a high-risk level would trigger an audio alert. The closer or denser a crowd is, the more risky it is considered. The crowd on the top left is considered the least risky and the crowd on the top right is considered the risky.

4.5 Crowd Risk Analysis

To evaluate the crowd-risk, we group all people into crowds based on physical proximity. Each person is represented as a point (x, y, d, v) in 4-dimensional space representing their 2D location, distance, and velocity. People are grouped into crowds with the same density-based clustering algorithm, DBSCAN [19], based on their proximity in the 4-dimensional representational space. The resulting crowd is represented by its size (number of people) s, average distance d, and motion v (signed real-number velocity). Then the mathematical formula below is used to evaluate the *riskiness* $r(\mathbf{c})$ of each crowd $\mathbf{c}$ (Detailed in Sup. Eqn. 9):

$$\begin{aligned} \mathbf{c} &\equiv s, d, v \\ r(\mathbf{c}) &= f(s, d, v) \end{aligned} \tag{3}$$

where $r : \mathbb{R}^3 \rightarrow \mathbb{R}^1 \in (0, 1)$ is a function that converts the features of a crowd into a real number representing the risk of the crowd; the crowd feature $\mathbf{c} \equiv s, d, v$ is a three-dimensional vector consisting of its size, distance, and motion. The overall riskiness $R(\mathbf{c})$ is the sum of individual riskiness of crowds:

$$R(\mathbf{c}) = \sum_{\mathbf{c} \in \mathbf{C}} r(\mathbf{c}) \tag{4}$$

where $\mathbf{C}$ is the set of all visible crowds. The individual crowd-risk function r is defined according to social distancing guidelines [7] such as considering 6 feet as the threshold for elevated risk (Sup. Sec. 2). Although other risk evaluation metrics that conform to the official advisory for different pandemics can also be selected. Once the overall crowd-risk is computed, our system sends different levels of audio-risk-alerts to the BVI user as shown in Fig. 5.

4.6 2D Audio Feedback

For every crowd deemed risk-inducing, our system sends a spatialized audio alert with the crowd-risk status (e.g. ”Moderate Risk”, ”High Risk”) so that the BVI know the level of risk associated with any nearby crowd and its general direction. The audio spatialization is implemented through open-sourced 3D sound-APIs like OpenSL, for mobile systems like Android [14] or through OpenAL [50] for Apple phones. The efficacy of spatial-directional audio in bone-conduction headphones has been well studied in [43] where participants were able to determine the direction of the audio source with bone-conduction

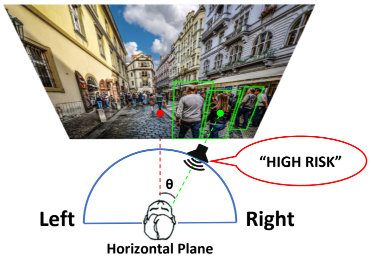


Fig. 6: 2D spatialized audio feedback system (Google Images CC License)

headphones on par with normal headphones. Studies have also shown that blind people are more sensitive to such binaural audio location cues than sighted people [48]. As demonstrated in Fig 6, after the cluster center of the crowd is computed, the offset from the image center to the cluster center is calculated in Euclidean metrics. To ensure a uniform spread of audio sources in-front of the user regardless of the input image size, we use the image previously re-scaled to scale of $[0, 100]$. Since we know the distance to the cluster center, we can calculate the center deviation angle $\theta \equiv \arcsin(\frac{cluster\_to\_img\_center\_offset}{distance\_to\_crowd})$, which is used by the spatial-sound API to generate a 2D directional audio. Finally, the severity of the crowd-risk is also conveyed through the intensity or volume of the audio alerts (i.e. higher risk equates to louder alerts).

5 Experimental Setup and Evaluation

Our tests are run in desktop environments (Sup. Sec. 6), however, our feature extraction module based on MobileNet-V2 [59] has real-time performances in smartphones. [31] experimentally showed that a quantized MobileNet model can run under 25ms at 40 fps with a Snapdragon 845 SoC enabled Qualcomm Hexagon Chipset smartphone. [31] and [39] also provide benchmarked object-detection performances for both Android and iOS smartphones based on other hardware chipsets as well.

5.1 Human Detection Performance

Dataset Setup: We train our human detection network on the PASCAL Visual Object Classes Challenge (VOC) 2007 and VOC 2012 dataset. The trained networks are tested on the PASCAL VOC 2007 test dataset. VOC 2007 and VOC 2012 have 9963 and 11540 images respectively with objects from over 20 different classes. Since we are concerned with human detection, we only report the overall mean average precision performance (mAP) and the results for the person class for object detection and localization task.

Table 1: Comparison of different models for human detection on the VOC 2007 test-set. All models were trained on the VOC 2007 and 2012 dataset. The frame rate is calculated for input images of size 500x375 pixels

Model	**mAP (Overall)**	**mAP (Person)**	**fps (GPU)**	**fps (CPU)**	**Model Size (MB)**
Our Network with SSDLite 320x320	0.7814	0.7403	**125±4**	**7±2**	**25.5**
VGG16 with SSD 300x300	0.9045	0.8751	60±5	3±2	201
Efficient B-Net with SSD 300x300	0.8166	0.7679	79±1	5±2	97.1

Results: Our proposed network, despite having a low mean average precision (mAP) for the people class detection, outperforms all other models in speed as frames per second (fps) and memory usage in both CPU and GPU settings. Our network sacrifices some accuracy for higher speed and lower memory usage, which is paramount if the network is used for detection in smartphones. The MobileNet-v2 backbone in our network has also been demonstrated to achieve a real-time performance of 40 fps in a GPU enabled-smartphone [31, 59] with model size reduced to 4.3MB.

5.2 Distance Regression Performance Analysis

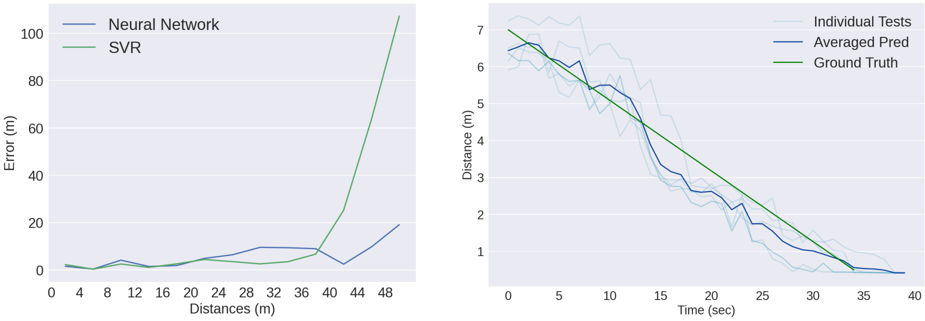
Dataset Setup: We train and evaluate our distance regressor networks on the KITTI Vision Benchmark Suite for 2D Object Detection [21], which has 7,481 training and 7,518 test images. Since we are only concerned with human-detection, we drop all classes except for the *Pedestrian* class. The *Cyclist* and *Person Sitting* classes are also dropped as their distributions deviate from that of the *Pedestrian* class when modeled using the bounding box heights and widths (Sup. Sec. 3 Fig. 1, Fig. 2). This leaves us with 1779 images and 4487 human instances. Since the test set is restricted to the KITTI Vision servers, we further divide the 1779 images training set into train and validation subsets with an $80 : 20$ split respectively. The images themselves are unnecessary for training the distance regressor as we only require the bounding box coordinates. We train the distance regressor network using the ADAM optimizer with a learning rate of 0.001 for 200 epochs with a batch size of 128. We also use a scheduler to reduce the learning rate to 10% of the previous rate if the validation loss plateaus for five epochs. The Support Vector Regressor (SVR) used for evaluation results in table 2 is based on the work done by [24]. The SVR is trained and evaluated on the same training and validation subset as the neural network regressor. The SVR is set up to use a radial basis function kernel with the hyper-parameter C set to 1.0 and epsilon set to 0.1.

Evaluation Metrics: Since our aim is to accurately predict the distances to the detected humans, we measure the performance of the distance regression models with metrics used in [16, 68], normally reserved for depth estimation. These metrics include the Mean Squared Error (MSE), the root of the mean squared error ($RMSE$), the log root mean squared error ($RMSE_{log}$), the absolute relative difference in distances ($Abs\ Rel$), and the squared relative difference in distances ($Squa\ Rel$) (Sup. Sec. 4).

Results: Fig. 7 and Table 2 demonstrate the effectiveness of our regression neural network compared to the SVR proposed in [24]. The experiments in Fig. 7b show the reliability of the distance predictions for a detected human approaching the camera. The SVR was found to be inaccurate for distances closer than 6 meters. For the experiment in Fig. 7b, the neural network regressor had a RMSE error of 0.4610 while the SVR had a RMSE error of 6.7508. However, Fig. 7a shows that both the SVR and the neural network regressor lose distance prediction performance as distances get longer since the detected bounding box shape gets smaller. Deterioration for distance estimation performance was also

Table 2: Distance prediction performance comparison for our validation subset split of the pedestrian class in the KITTI–object-detection dataset

Regressor	lower is better				
	MSE	**RMSE**	**RMSE$_{log}$**	**Squa Rel**	**Abs Rel**
Neural Network Regressor	6.0088	2.4513	0.1304	0.3911	0.0788
Support Vector Regressor	7.0425	2.6538	0.1217	0.3215	0.0782



(a) MSE over different distances for Neural Network and SVR.

(b) Distance predicted by Neural Network compared to ground truth for a human approaching the camera

Fig. 7: Distance Regressor Performance

observed in [64, 64, 68]. This is not a concern for us since we put more emphasis on how the models perform for short distances within 7 meters where social distancing guidelines might come into action.

5.3 Crowd Motion Tracking Performance Analysis

Dataset Setup: To evaluate our crowd-motion analysis setup, we make use of the 2016 Multiple-Object-Tracking Benchmark [47] (*MOT16*) training subset. We use the *MOT16*-02 subset of the *MOT16* data to evaluate different human detection backbone networks as reported in Table 3. We only consider the pedestrian class with a visibility ratio of more than 0.7 in this subset for evaluation and drop all non-pedestrian classes from our test set. We use the *MOT 1602* train subset as the testing subsets are only available for server evaluation on all classes explicitly. This subset contains a video sequence of 600 images of $1920x1080$ resolution. The *MOT* benchmarks we report in table 3 are based on this subset.

Evaluation Metrics: Our primary objective in crowd-motion analysis is to measure the motion (changing distance) of crowds and individuals from the BVI user. To accurately calculate this motion, we need to track the detected humans. Therefore, we use the *CL*assification of *E*vents, *A*ctivities, and *R*elationships (*CLEAR*) metrics [1] for the task of Multiple Object Tracking [47] (*MOT*) to

Table 3: Comparison of the object tracking metrics on the pedestrian class of the *MOT 2016-02 subset* given backbone networks and SSD detectors of different input sizes. (Detection score threshold was set to 0.65 for all models)

Model	higher is better			lower is better			
	MOTA(%)	**MOTP(%)**	**MT(%)**	**ML(%)**	**FP**	**FN**	**IDsW**
Our Network with SSDLite 320x320	**8.1**	**28.9**	4.65	76.74	**377**	**3378**	**12**
VGG with SSD 300x300	4.5	27.8	5.26	73.68	950	3302	29
Efficient Net b3 with SSD 300x300	7.3	28.6	5.26	81.57	563	3495	13

benchmark the human-motion tracking performance. The *CLEAR* metrics in Table 1 include standardized metrics such as the Multiple Object Tracking Accuracy (*MOTA*) and the Multiple Object Tracking Precision (*MOTP*) (Sup. Sec. 5). Additional tracking metrics also include the Mostly Tracked (*MT*), Mostly Lost(*ML*), False Positives (*FP*), False Negatives(*FN*), and Identity switches (*IDsW*) (Sup. Sec. 5).
Results: Table 3 shows that our backbone network is the most suitable model for the human tracking task required to create the motion maps in Fig. 4 as our network has the highest *MOTA* and *MOTP*. However, these metric scores should not be alarming since the *MOT16* benchmark contains excessive detectable objects in each frame with frequently occluded paths while even state of the art methods only achieved around 33.7 % *MOTA* in the *MOT16* testset [47].

6 Conclusion

Given that previous research and commercial solutions designed to aid the BVI community do not take health risks associated with a pandemic into account, in this paper we have presented and demonstrated the efficacy of an active crowd analysis system to help mitigate these pandemic-related health risks for the BVI. Our smartphone-based system combines crowd density, distance, and motion analysis to detect the risks associated with nearby crowds and relay this risk to the BVI individual through a directional 2D audio. Active Crowd Analysis, in aggregate, enables such sight-impaired persons to maintain a safe physical distance from other humans or crowds meeting the official social distancing guidelines to avoid the spread of contagions during a pandemic.

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. Bhowmick, A., Hazarika, S.M.: An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends. Journal on Multimodal User Interfaces **11**(2), 149–172 (2017)
3. BlindSquare: Pioneering accessible navigation – indoors and outdoors (May 2020), https://www.blindsquare.com/
4. Bologna, G., Deville, B., Pun, T., Vinckenbosch, M.: Transforming 3d coloured pixels into musical instrument notes for vision substitution applications. EURASIP Journal on Image and Video Processing **2007**, 1–14 (2007)
5. Bourne, R.R., Stevens, G.A., White, R.A., Smith, J.L., Flaxman, S.R., Price, H., Jonas, J.B., Keeffe, J., Leasher, J., Naidoo, K., et al.: Causes of vision loss worldwide, 1990–2010: a systematic analysis. The Lancet Global Health **1**(6), e339–e349 (2013)
6. Brainport: Disabilities technology: Brainport technologies: United states, https://www.wicab.com/
7. CDC: Social distancing, quarantine, and isolation (May 2020), https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html
8. Chahyati, D., Fanany, M.I., Arymurthy, A.M.: Tracking people by detection using cnn features. Procedia Computer Science **124**, 167–172 (2017). https://doi.org/10.1016/j.procs.2017.12.143
9. Coniglio, C., Meurie, C., Lézoray, O., Berbineau, M.: People silhouette extraction from people detection bounding boxes in images. Pattern Recognition Letters **93**, 182–191 (Jul 2017). https://doi.org/10.1016/j.patrec.2016.12.014
10. Corp, A.T.: Flattening the inaccessibility curve, https://flatteninaccessibility.com/
11. Coughlan, J.M., Miele, J.: Ar4vi: Ar as an accessibility tool for people with visual impairments. In: 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct). pp. 288–292. IEEE (2017)
12. Dakopoulos, D., Bourbakis, N.G.: Wearable obstacle avoidance electronic travel aids for blind: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **40**(1), 25–35 (2009)
13. Dale, Z.: Experiences of deafblind persons during the covid-19 outbreak. International Disability Alliance (2020), http://www.internationaldisabilityalliance.org/content/experiences-deafblind-amid-covid-19-outbreak
14. Deveopers, A.: Android ndk, https://developer.android.com/ndk/guides/audio
15. Dewhurst, D.C.: Audiotactile vision substitution system (Aug 7 2012), uS Patent 8,239,032
16. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. arXiv:1406.2283 [cs] (Jun 2014), http://arxiv.org/abs/1406.2283, arXiv: 1406.2283
17. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2155–2162 (2014)

18. Ess, A., Leibe, B., Van Gool, L.: Depth and appearance for mobile scene analysis. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8 (2007)
19. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. p. 226–231. KDD'96, AAAI Press (1996)
20. Fox, D., Kar, R., Li, A., Pandey, A.: Augmented reality for visually impaired people, `https://www.ischool.berkeley.edu/projects/2019/augmented-reality-visually-impaired-people`
21. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
22. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. p. 580–587. IEEE (Jun 2014). https://doi.org/10.1109/CVPR.2014.81
23. Giudice, N.A., Legge, G.E.: Blind Navigation and the Role of Technology, chap. 25, pp. 479–500. John Wiley & Sons, Ltd (2008). https://doi.org/10.1002/9780470379424.ch25
24. Gökçe, F., Üçoluk, G., Sahin, E., Kalkan, S.: Vision-based detection and distance estimation of micro unmanned aerial vehicles. Sensors (Basel, Switzerland) **15**, 23805 – 23846 (2015)
25. Google: Google glass, `https://www.google.com/glass/start/`
26. Gordon, K.D.: survey: The impact of the covid-19 pandemic on canadians who are blind deaf-blind, and partially-sighted (2020), `http://ccbnational.net/shaggy/wp-content/uploads/2020/05/COVID-19-Survey-Report-Final-wb.pdf`
27. Haseeb, M., Guan, J., Ristić-Durrant, D., Gräser, A.: Disnet: A novel method for distance estimation from monocular camera. In: IEEE/RSJ Internation Conference on Intelligent Robots and Systems (IROS, Spain 2018). ieee (2018)
28. Haymes, S.A., Johnston, A.W., Heyes, A.D.: Relationship between vision impairment and ability to perform activities of daily living. Ophthalmic and Physiological Optics **22**(2), 79–91 (2002)
29. Helal, A., Moore, S.E., Ramachandran, B.: Drishti: An integrated navigation system for visually impaired and disabled. In: Proceedings fifth international symposium on wearable computers. pp. 149–156. IEEE (2001)
30. HTC: Htc vive, `https://www.vive.com/us/product/`
31. 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
32. IrisVision: Wearable low vision glasses for visually impaired (May 2020), `https://irisvision.com/`
33. Jackson, A.: The hidden struggles america's disabled are facing during the coronavirus pandemic. CNBC News (2020), `https://www.cnbc.com/2020/05/10/the-struggles-americas-disabled-are-facing-during-coronavirus-pandemic.html`
34. Kajabad, E.N., Ivanov, S.V.: People detection and finding attractive areas by the use of movement detection analysis and deep learning approach. Procedia Computer Science **156**, 327–337 (2019). https://doi.org/10.1016/j.procs.2019.08.209
35. Kajimoto, H., Kanno, Y., Tachi, S.: Forehead electro-tactile display for vision substitution. In: Proc. EuroHaptics (2006)

36. Kempen, G.I., Ballemans, J., Ranchor, A.V., van Rens, G.H., Zijlstra, G.R.: The impact of low vision on activities of daily living, symptoms of depression, feelings of anxiety and social support in community-living older adults seeking vision rehabilitation services. Quality of life research **21**(8), 1405–1411 (2012)
37. Kulyukin, V., Gharpure, C., Nicholson, J., Pavithran, S.: Rfid in robot-assisted indoor navigation for the visually impaired. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566). vol. 2, pp. 1979–1984. IEEE (2004)
38. Lamoureux, E.L., Hassell, J.B., Keeffe, J.E.: The determinants of participation in activities of daily living in people with impaired vision. American journal of ophthalmology **137**(2), 265–270 (2004)
39. Lee, J., Chirkov, N., Ignasheva, E., Pisarchyk, Y., Shieh, M., Riccardi, F., Sarokin, R., Kulik, A., Grundmann, M.: On-Device Neural Net Inference with Mobile GPUs. arXiv:1907.01989 [cs, stat] (Jul 2019), http://arxiv.org/abs/1907.01989, arXiv: 1907.01989
40. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. Lecture Notes in Computer Science p. 21–37 (2016)
41. Loomis, J.M., Golledge, R.G., Klatzky, R.L.: Gps-based navigation systems for the visually impaired. Fundamentals of wearable computers and augmented reality **429**, 46 (2001)
42. Loomis, J.M., Golledge, R.G., Klatzky, R.L., Speigle, J.M., Tietz, J.: Personal guidance system for the visually impaired. In: Proceedings of the first annual ACM conference on Assistive technologies. pp. 85–91 (1994)
43. MacDonald, J.A., Henry, P.P., Letowski, T.R.: Spatial audio through a bone conduction interface: Audición espacial a través de una interfase de conducción ósea. International Journal of Audiology **45**(10), 595–599 (Jan 2006). https://doi.org/10.1080/14992020600876519
44. Maidenbaum, S., Abboud, S., Amedi, A.: Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation. Neuroscience & Biobehavioral Reviews **41**, 3–15 (2014)
45. Microsoft: Microsoft hololens, https://www.microsoft.com/en-us/hololens
46. Microsoft: Seeing ai, https://www.microsoft.com/en-us/ai/seeing-ai
47. 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
48. Nilsson, M.E., Schenkman, B.N.: Blind people are more sensitive than sighted people to binaural sound-location cues, particularly inter-aural level differences. Hearing Research **332**, 223–232 (Feb 2016). https://doi.org/10.1016/j.heares.2015.09.012, https://linkinghub.elsevier.com/retrieve/pii/S0378595515300174
49. Okonji, P.E., Ogwezzy, D.C.: Awareness and barriers to adoption of assistive technologies among visually impaired people in nigeria. Assistive Technology **31**(4), 209–219 (2019)
50. OpenAL: Open audio library (2020), https://openal.org/, [Accessed 20-July-2020]
51. Praveen, S.: Efficient depth estimation using sparse stereo-vision with other perception techniques. In: Radhakrishnan, S., Sarfraz, M. (eds.) Coding Theory, chap. 7. IntechOpen, Rijeka (2020). https://doi.org/10.5772/intechopen.86303
52. Qiu, S., Han, T., Osawa, H., Rauterberg, M., Hu, J.: Hci design for people with visual disability in social interaction. In: International Conference on Distributed, Ambient, and Pervasive Interactions. pp. 124–134. Springer (2018)

53. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)
54. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv:1804.02767 [cs] (Apr 2018), `http://arxiv.org/abs/1804.02767`, arXiv: 1804.02767
55. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
56. Rezaei, M., Terauchi, M., Klette, R.: Robust vehicle detection and distance estimation under challenging lighting conditions. IEEE Transactions on Intelligent Transportation Systems **16**(5), 2723–2743 (2015)
57. Roentgen, U.R., Gelderblom, G.J., Soede, M., De Witte, L.P.: Inventory of electronic mobility aids for persons with visual impairments: A literature review. Journal of Visual Impairment & Blindness **102**(11), 702–724 (2008)
58. S., K., S., V.: Contour-based object tracking in video scenes through optical flow and gabor features. Optik **157**, 787–797 (Mar 2018). https://doi.org/10.1016/j.ijleo.2017.11.181
59. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks (2018)
60. 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
61. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs] (Apr 2015), `http://arxiv.org/abs/1409.1556`, arXiv: 1409.1556
62. Singh, P., Deepak, B., Sethi, T., Murthy, M.D.P.: Real-time object detection and tracking using color feature and motion. In: 2015 International Conference on Communications and Signal Processing (ICCSP). p. 1236–1241. IEEE (Apr 2015). https://doi.org/10.1109/ICCSP.2015.7322705
63. Stelmack, J.: Quality of life of low-vision patients and outcomes of low-vision rehabilitation. Optometry and Vision Science **78**(5), 335–342 (2001)
64. Tuohy, S., O'Cualain, D., Jones, E., Glavin, M.: Distance determination for an automobile environment using inverse perspective mapping in opencv. In: IET Irish Signals and Systems Conference (ISSC 2010). p. 100–105. IET (2010). https://doi.org/10.1049/cp.2010.0495
65. Valve: Valve index, `https://store.steampowered.com/valveindex`
66. WHO: Global data on visual impairment 2010 (2010), `https://www.who.int/blindness/GLOBALDATAFINALforweb.pdf`
67. Zhang, R., Li, Y., Zhang, A.L., Wang, Y., Molina, M.J.: Identifying airborne transmission as the dominant route for the spread of covid-19. Proceedings of the National Academy of Sciences **117**(26), 14857–14863 (2020). https://doi.org/10.1073/pnas.2009637117
68. Zhu, J., Fang, Y.: Learning Object-Specific Distance From a Monocular Image. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3838–3847. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00394