Machine Learning with Applications 10 (2022) 100427

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Automatic social distance estimation for photographic studies: Performance evaluation, test benchmark, and algorithm ☆

Mert Seker [a], Anssi Männistö [b], Alexandros Iosifidis [c,1], Jenni Raitoharju [d,1,*]

[a] Unit of Computing Sciences, Tampere University, Tampere, Finland
[b] Unit of Communication Sciences, Tampere University, Tampere, Finland
[c] Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark
[d] Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland

ARTICLE INFO

Keywords:
Social distance estimation
Person detection
Human pose estimation
Performance evaluation
Test benchmark
Proxemics

ABSTRACT

The social distancing regulations introduced to slow down the spread of COVID-19 virus directly affect a basic form of non-verbal communication, and there may be longer term impacts on human behavior and culture that remain to be analyzed in proxemics studies. To obtain quantitative results for such studies, large media and/or personal photo collections must be analyzed. Several social distance monitoring methods have been proposed for safety purposes, but they are not directly applicable to general photo collections with large variations in the imaging setup. In such studies, the interest shifts from safety to analyzing subtle differences in social distances. Currently, there is no suitable benchmark for developing such algorithms. Collecting images with measured ground-truth pair-wise distances using different camera settings is cumbersome. Moreover, performance evaluation for these algorithms is not straightforward, and there is no widely accepted evaluation protocol. In this paper, we provide an image dataset with measured pair-wise social distances under different camera positions and settings. We suggest a performance evaluation protocol and provide a benchmark to easily evaluate such algorithms. We also propose an automatic social distance estimation method that can be applied on general photo collections. Our method is a hybrid method that combines deep learning-based object detection and human pose estimation with projective geometry. The method can be applied on uncalibrated single images with known focal length and sensor size. The results on our benchmark are encouraging with 91% human detection rate and only 38.24% average relative distance estimation error among the detected people.

1. Introduction

Social distances are a part of non-verbal human communications and, naturally, there are personal and cultural differences in how people feel about their personal space and interpret the interpersonal distance in different situations. The research field under social studies concerning these phenomena related to space is known as *proxemics* (Hall et al., 1968). Despite the long history of studies in the field (Cook, 1970; Hall, 1966; Harrigan, 2005), it remains difficult to carry out quantitative analysis on the actual social distances in the natural situations outside of monitored test conditions, e.g., when people are spending their free time with their families. One way to approach this problem is *visual social distancing* (VSD), where the interpersonal distances are automatically measured from the images or videos. A comprehensive overview of the VSD problem, including the main challenges and connections to social studies, is provided in Cristani, Bue, Murino, Setti, and Vinciarelli (2020).

Social distancing has received a lot of attention in recent years due to the outbreak of SARS-CoV-2, also known as COVID-19, virus that was declared as a global pandemic by the World Health Organization (WHO) in March 2020. Social distancing played an important role in slowing down the spread of the virus and WHO recommended to stay at least one meter apart from other people in order to reduce the risk of infection. Automatically monitoring the social distances during pandemic restrictions is important for safety reasons, but it is also interesting to analyze how the restrictions globally changed basic human behavior (Di Corrado et al., 2020; Eden, Johnson, Reinecke, & Grady, 2020; Zhang, Gao, Gross, Shrum, & Hayne, 2021). After the

☆ M. Seker, A. Männistö, and J. Raitoharju would like to acknowledge the financial support from Helsingin Sanomat foundation, project "Machine learning based analysis of the photographs of the corona crisis". A. Iosifidis acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 957337 (MARVEL).

* Corresponding author.
E-mail addresses: mert.seker@tuni.fi (M. Seker), anssi.mannisto@tuni.fi (A. Männistö), ai@ece.au.dk (A. Iosifidis), jenni.k.raitoharju@jyu.fi (J. Raitoharju).
[1] EURASIP member.

https://doi.org/10.1016/j.mlwa.2022.100427
Received 26 May 2022; Received in revised form 13 October 2022; Accepted 18 October 2022
Available online 9 November 2022

Fig. 1. An example of an image that represents a style, which is common in personal and media photography, but not in monitoring. The image is included in our test set.

pandemic, there are many interesting research questions in proxemics and other fields to look into: how the social distancing affected everyday life, what kind of significant differences were there between different countries, can the differences be linked to the spreading speed, will there be any long-term changes that will stay after the pandemic.

While there are methods and sensors available for automatic monitoring of social distances (Nguyen et al., 2020), the analysis of deeper and longer term social and cultural impacts of the social distancing regulations requires looking into different source data, such as personal photo collections and pictures published in newspapers and magazines. For monitoring purposes, it is possible to use fixed camera setup and location, take videos or simultaneous images from multiple viewpoints, and use additional sensors such as depth or thermal cameras. All these can make the social distance estimates more accurate but are not available for typical personal and media photos that are not taken with a fixed setup, but have varying parameters such as focal length, sensor size, lighting conditions, and pitch angle. An example of an image that could be found in a personal or media photo collection, but not in a monitoring or surveillance setup is shown in Fig. 1. At the same time, in social and proxemics studies the focus shifts from monitoring whether people are obeying the regulations to more subtle differences in the social distances and how they are represented in the media.

During the pandemic, most effort has been understandably on the monitoring side, and currently there is no suitable benchmark for developing and testing algorithms for accurate social distance analysis from single images having varying camera parameters. This can be due to the laboriousness of gathering varying images with measured pair-wise distances between humans. At the same time, there is no clear protocol for measuring the algorithm performance in this task. To address these lacks, we provide a social distance evaluation test benchmark including a protocol for mapping the detected pair-wise distances into the corresponding ground truth distances, a suggested overall performance metric, and 300 test images taken with varying setups: indoors–outdoors, sitting–standing, varying camera angles using 2 different cameras and 7 different focal lengths. The photos were taken by a professional photojournalist to follow the typical media photography style. We publish also easy-to-use codes for evaluating novel methods and make it easy to integrate additional test photos.

We also propose a social distance estimation algorithm that can be applied on any uncalibrated single image taken by a regular camera as long as focal length and sensor size are known. It is a hybrid method that combines deep learning-based object detection and human pose estimation with projective geometry using image parameters (focal length, sensor size) and pixel locations. While the results are promising, we also point out some of the main remaining challenges for future development.

The rest of the paper is organized as follows. Section 2 introduces related work on social distancing and automatic distance evaluation. Section 3 describes the provided test benchmark and the proposed evaluation protocol. Our method for automatic social distance estimation is described in Section 4. Section 5 provides our experimental setup and results and, finally, Section 6 concludes the paper.

2. Related work

Effectiveness of social distancing on slowing down the spread of the COVID-19 virus has been widely studied (Abouk & Heydari, 2021; Balasa, 2020; Courtemanche, Garuccio, Le, Pinkston, & Yelowitz, 2020; Prem et al., 2020; Sun & Zhai, 2020; Vokó & Pitter, 2020), and these studies confirm that social distancing measures are successful in reducing the growth rate of the virus. Therefore, monitoring and regulating the social distancing behavior between people has played a crucial part in dampening the effects of the virus. In addition to directly effecting the virus spread, social distancing has globally changed human behavior and interactions leading to different side-impacts, e.g., on mental health (Ford, 2020; Jacob et al., 2020), physical activity (Di Corrado et al., 2020; Jacob et al., 2020), mood and memory (Zhang et al., 2021), and media consumption (Eden et al., 2020). Such impacts and their cross-cultural (Al-Hasan, Khuntia & Yim, 2020; Al-Hasan, Yim & Khuntia, 2020; Doogan, Buntine, Linger, & Brunt, 2020) and cross-sectional (Jacob et al., 2020; Lee, Kang, & You, 2021) differences continue to draw attention from researchers in many fields.

Social distance monitoring for safety reasons can be eased by automatic social distance estimation from images and videos. A comprehensive survey in Nguyen et al. (2020) explores the wide array of current technologies that can be used to monitor and encourage social distancing. A commercial pedestrian tracking system was used in Pouw, Toschi, van Schadewijk, and Corbetta (2020) to detect passengers in crowded environments and estimate the distances between them by using a graph based approach. A study in Ahmed, Ahmad, Rodrigues, Jeon, and Din (2021) proposed using a deep learning based model with YOLOv3 (Redmon & Farhadi, 2018) as its backbone to monitor social distancing violations from overhead view cameras. In Punn, Sonbhadra, and Agarwal (2020), the authors used YOLOv3 and DeepSort (Wojke & Bewley, 2018; Wojke, Bewley, & Paulus, 2017) to detect bounding boxes of people in RGB images and by utilizing these bounding boxes, they detected the cases of social distance violations.

A work in Aghaei et al. (2021) proposed to use skeleton keypoints generated from human body pose estimation algorithms (Cao, Hidalgo Martinez, Simon, Wei, & Sheikh, 2019) to estimate the distance between people from uncalibrated images. The authors used manual tuning to estimate the homography matrix (Young, 1982) of an image plane and then used leg, arm, and torso lengths of the people alongside with the homography matrix to draw a safe space circle underneath every detected person. Then, any collision between the estimated safe space circles was reported as a social distance violation. Similarly, the work in Fabbri et al. (2020) takes advantage of manual homography matrix calibration to estimate social distances for fixed cameras. Separating the work from Aghaei et al. (2021), bounding boxes obtained from the object detection model (Zhou, Wang, & Krähenbühl, 2019) and the height of these boxes are used as reference points to estimate the locations of the people. Moreover, a small CNN is used to estimate the feet locations even when they are not visible. The output of this CNN is used to correct the height of the bounding boxes in cases of occlusions. Another similar study in Yang, Yurtsever, Renganathan, Redmill, and Özgüner (2020) also used bounding boxes obtained from object detectors (Bochkovskiy, Wang, & Liao, 2020; Ren, He, Girshick, & Sun, 2016) to estimate locations of the people from surveillance camera footage by using the homography matrix that is calculated from the known extrinsics.

The work in Bertoni, Kreiss, and Alahi (2021) used a feed forward neural network that was trained on the intrinsic parameters of the

camera and the keypoints obtained from a pose estimation model. The model outputs the predicted 3D locations as well as the orientations of the detected people. While detecting safe distance violations, not only the proximity but also the orientation of the people with respect to one another is considered. Finally, the study in Morerio, Bustreo, Wang, and Bue (2021) proposed a neural network architecture that takes a pair of 2D body keypoints as input and outputs the estimated pair-wise distance. The two sets of body keypoints are converted into feature vectors by an encoder block. The vectors are then concatenated and given as input to a regressor block, followed by a fully connected layer that was trained on the public datasets Epfl-Mpv-VSD (Fleuret, Berclaz, Lengagne, & Fua, 2008), Epfl-Wildtrack-VSD (Chavdarova et al., 2018), OxTown-VSD (Benfold & Reid, 2011) and Kitti (Geiger, Lenz, & Urtasun, 2012) to estimate pair-wise distances. The output of the regressor block is also used as input to another branch with a gradient reversal layer (Ganin et al., 2016) to estimate the camera's tilt angle and height from the ground plane in order to make the estimations more robust to variations in camera viewpoints. The method works on any single uncalibrated image.

Most of the introduced works approach automatic social distance estimation as a monitoring or surveillance task, where the goal is to prevent social distance regulation violations. To this end, they apply additional sensors, use predefined camera settings, and/or manually define a homography matrix for a certain environment. While such approaches can improve the social distance estimation accuracy, they are not feasible when the purpose is to analyze the impacts of social distances from personal or media photo collections.

Moreover, the above-mentioned studies approach the automatic social distance estimation problem as a binary classification problem where they aim to classify the pair-wise distances between people either as safe or unsafe, depending on a given threshold. Classifying distances in a binary manner has a high tolerance for distance estimation errors. For example, if the threshold for safe distance is set to 2 m, the actual distance between a pair of people is 1.9 m, and a method estimates that distance as 0.1 m, the percentual distance estimation error would be 94.7%, but a binary classification approach would still correctly label the situation as a social distance violation. Furthermore, the binary approach does not provide any additional information on the severity of the violations in different situations which may be relevant information for subsequent analysis.

A common pattern observed in most of the machine learning based social distance estimation methods (with the exception of at least (Aghaei et al., 2021; Bertoni et al., 2021; Morerio et al., 2021) that use keypoints of the human body) is that they rely on the bounding boxes drawn by object detectors to detect social distance violations. Although the current object detectors are accurate in detecting objects, the bounding boxes are generally loosely drawn around these objects. Thus, it is not reliable to use only the bounding box information for estimating exact distances between people as it is not possible to infer accurate 3D location estimates from the bounding boxes alone. Therefore, we aim to estimate exact 3D locations of all the people in uncalibrated RGB images with respect to the camera by using the information extracted from the human body skeleton detected by body estimation algorithms. Moreover, we also incorporate an object detection model for people detection. However, the purpose of the people detection in our approach is to only detect the false positives in skeleton keypoints, when they are drawn on non-human objects.

The method in Aghaei et al. (2021) is the most similar to our method as it also uses body poses. In Aghaei et al. (2021), manual input is used to estimate the homography matrix of the image plane to the ground plane. The method is evaluated on surveillance camera footage and the task is approached as a binary classification problem. It is feasible to manually set the homography matrix of surveillance cameras as these cameras are generally non-moving and stable. Contrary to this, we want our method to be fully automatic as we aim to estimate distances in images taken in different locations with different cameras. Instead of requiring manual input to estimate the homography as the study in Aghaei et al. (2021), we assume that we can find keypoint pairs that are parallel to camera's sensor plane and we use the image parameters, i.e., focal length and sensor size in our distance estimation.

For the developing and testing social distance estimation methods, it is important to have image datasets that have a suitable setup and ground-truth for the task. The previous works have used datasets such as Epfl-Mpv-VSD, Epfl-Wildtrack-VSD and OxTown-VSD. These datasets include videos taken by surveillance cameras with fixed extrinsic and intrinsics and they do not include manually measured ground truth locations and distances. Instead, the locations of the people are estimated by making use of the annotation boxes that were drawn on the people. The pixel locations of these annotation boxes are used as a reference point to estimate the subjects' locations by taking the extrinsic parameters into account. This means that these locations are not exactly ground truth, but estimations based on the known extrinsics and the pixel locations of the manually annotated person bounding boxes. Furthermore, since exact body parts are not annotated and the annotations are only in bounding box format, it is not feasible nor possible to accurately match the detected people with the given ground truth people when there are multiple overlapping boxes. Moreover, only the people that are passing on a certain region of interest are annotated.

Due to the aforementioned reasons, the existing datasets are not suitable for evaluating methods that aim at estimating distances in general photo collections and are not manually tuned for a specific camera and environments. Furthermore, the approximate person annotations and location estimates do not allow accurately measuring the distance estimation performance, but are only suitable for detecting coarse violations in social distancing recommendations. While this may be sufficient for surveillance purposes in fixed environments, more accurate ground-truth and annotations are needed for evaluating methods aiming at detecting subtle changes in long-term social distancing behavior in varying environments. In the following section, we introduce our novel dataset that addresses the mentioned drawbacks of the existing datasets.

3. KORTE social distance estimation benchmark

We provide a test benchmark for facilitating research in automatic social distance evaluation. We propose a performance evaluation protocol and provide 300 test images with ground-truth pair-wise distances. While the number of images is too low for training fully learning-based systems, it provides a varied test setup. All the evaluation codes along with the test photos are publicly available at https://doi.org/10.23729/b2ea87e6-b845-46b8-abf3-cdbe299ce8b0. It is also easy to complement the benchmark with additional images by following the proposed annotation format and using the provided evaluation protocol.

3.1. Test photo collection

We collected test photos in four separate photo shoots. The first and third photo shoots were organized outdoors at Tampere University campus in December 2020 and August 2021, respectively. Every person was standing. The second and fourth photo shoots were organized indoors at Tampere University campus in January 2021 and August 2021 with people sitting around tables and sofas. We had 6 volunteer test subjects in the first and second photo shoots and 7 volunteer test subjects in the third and fourth photo shoots. We followed the COVID-19 restrictions at the time: everyone was wearing a mask and we were less than 10 people gathering. As an additional safety measure, we placed to closest distances from each other only people who meet regularly anyway because they share working space or live together. Every test subject signed an agreement allowing to use their images for research purposes. Any bypassers in the images were censored out to

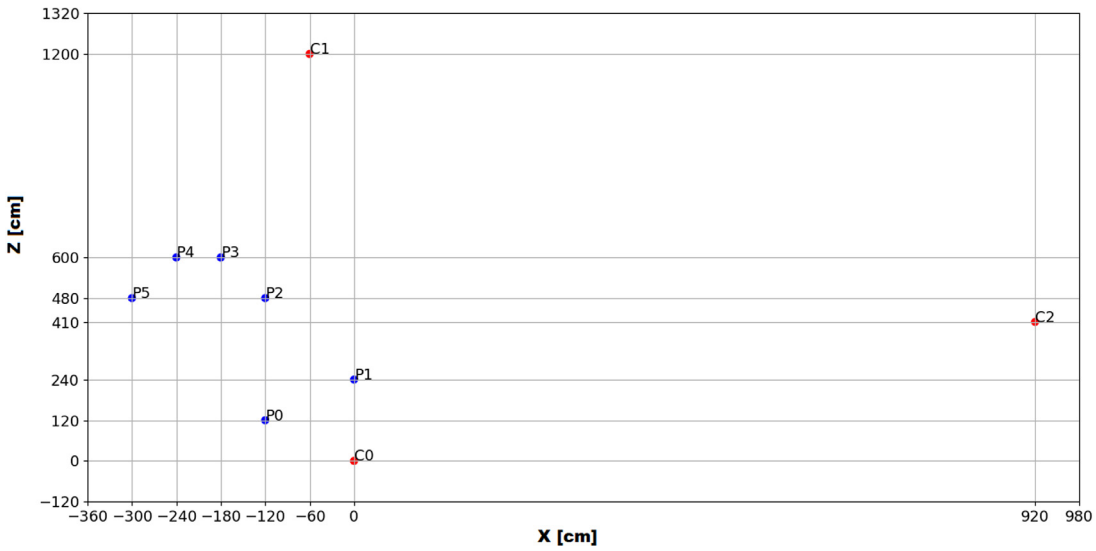

Fig. 2. Birdseye view of the first photo shoot (outdoor). The ground truth locations of the people and cameras are given in blue and red dots, respectively.

Fig. 3. Example photos from the test dataset. The upper row has photos from the first photo shoot (outdoor) taken from all camera positions C0 (left) to C2 (right) and the lower row has photos from the second photo shoot (indoor) taken from all camera positions C0 (left) to C3 (right).

respect their privacy and because their exact positions were unknown. The photos were taken by a professional photojournalist.

During the photo shoots, test subjects stayed on the same known positions, while the photographer changed his position and used multiple cameras and lenses at each spot. Fig. 2 shows as an example the birdseye view of the first photo shoot (outdoor). P0, P1, P2, P3, P4, P5 are the locations of the 6 test subjects and C0, C1, C2 are the camera locations. For the first photo shoot, P0, P1, P2, P3, P4, P5, C0 and C1 were all on the same ground plane, while C2 was at a balcony with a height of 230 cm relative to the ground plane that all of the other locations were at. Similar birdseye views of the other photo shoots are included in the Appendix A.12–A.14. The unit of the x and z axis labels is centimeters. The ground truth locations of the cameras and the test subjects were measured and maintained exploiting tiles on the ground/floor that were equal in size. While test subjects' locations were fixed during each photo shoot, they were asked to vary their orientation and pose. The ground truth locations of all the cameras and test subjects for all the photo shoots are provided with the dataset.

We do not report the exact pitch angles, and they were not fixed in the photo shoots. Due to the camera positions, pitch angles are close to zero in most of the images except for the 54 photos taken from camera position C2 in the first and third photo shoot, where the camera was at an elevated position. We believe that our dataset represents a typical media or personal photo collection with respect to the pitch angles, but it should be noted that methods performing well on our dataset (especially if they rely on the zero pitch angle assumption) may not perform equally well on extreme pitch angles, such as overhead images.

The used camera models were Canon EOS 5D Mark II and Canon EOS 6D Mark II. The used focal lengths were 16, 24, 35, 50, 105, 200, and 300 mm. The cameras were stabilized on a tripod. Fig. 3 shows example photos from the first and second photo shoots, one photo from each camera position.

3.2. Test data description

The overall dataset contains 300 images including 174 outdoor images and 126 indoor images. All of the images are in JPG format. The resolutions of the images are 2400 × 1600, 4080 × 2720 and 4160 × 2768 with 139, 80 and 81 images in each resolution, respectively. Two different camera models were used and the sensor size for both of these cameras is 36 mm in width and 24 mm in height. The distribution of the pictures in terms of focal lengths, camera models, and shooting settings is given in Table 1.

Along with the images, we also provide different annotation data provided in three separate .csv files illustrated in Fig. 4. The first file

Person Tag	Body Part	Pixel X Pos.	Pixel Y Pos.	Filename	Image Width	Image Height
P1	Eyes	1381	1281	IMG_6285.jpg	4160	2768
P1	Head	1359	1287	IMG_6285.jpg	4160	2768
P1	Shoulder	1338	1344	IMG_6285.jpg	4160	2768
P0	Eyes	1574	1274	IMG_6285.jpg	4160	2768
P0	Head	1545	1287	IMG_6285.jpg	4160	2768
P0	Shoulder	1516	1344	IMG_6285.jpg	4160	2768
P4	Shoulder	2268	1326	IMG_6285.jpg	4160	2768
P4	Head	2286	1277	IMG_6285.jpg	4160	2768
P3	Eyes	2375	1283	IMG_6285.jpg	4160	2768
P3	Head	2385	1292	IMG_6285.jpg	4160	2768

(a) Body part pixel locations

Photo shoot ID	Person or Camera Tag	x	y	z
0	C1	-600	0	12000
0	C2	9200	0	4100
0	P0	-1200	0	1200
0	P1	0	0	2400

Unit is mm

(b) Ground truth relative 3D location

Photo shoot ID	Camera Locations	Filename
1	C2	IMG_6285.jpg
1	C2	IMG_6286.jpg
1	C2	IMG_6289.jpg
1	C2	IMG_6292.jpg
1	C2	IMG_6295.jpg
1	C2	IMG_6296.jpg
1	C2	IMG_6297.jpg
1	C2	IMG_6298.jpg
1	C3	IMG_6302.jpg

(c) Photo shoot identifiers and camera locations

Fig. 4. Annotation file formats.

Table 1
Numbers of photos in the test dataset for different focal lengths (mm), camera models, and shooting settings (indoor/outdoor).

Focal length	Camera model	Shooting setting	
	Canon	Indoor	Outdoor
16	EOS 5D Mark II	–	6
16	EOS 6D Mark II	12	10
24	EOS 5D Mark II	5	8
24	EOS 6D Mark II	25	8
35	EOS 5D Mark II	–	–
35	EOS 6D Mark II	24	24
50	EOS 5D Mark II	–	31
50	EOS 6D Mark II	23	30
105	EOS 5D Mark II	15	–
105	EOS 6D Mark II	22	32
200	EOS 5D Mark II	–	7
200	EOS 6D Mark II	–	–
300	EOS 5D Mark II	–	8
300	EOS 6D Mark II	–	10
All		126	174

(Fig. 4(a)) contains the pixel locations of four different body parts. These annotated body parts are the center of the eyes, the center of the shoulders, the center of the torso, and the center of the head. If a body part is not visible in the image, it is not annotated. The people in the images are labeled as P0, P1, P2, P3, P4, P5, P6, P7, and P8 in the annotation file. These person tags are consistent through all of the images. This means that a person tag always refers to the same person in all of the images that we provide. The second file (Fig. 4(b)) contains the 3D locations of people and different camera positions in all photo shoots. Photo shoot IDs 0, 1, 2, and 3 refer to the first (outdoor), second (indoor), third (outdoor), and fourth (indoor) photo shoots, respectively. The third file (Fig. 4(b)) links the image filenames with the corresponding photo shoot and camera location. The cameras' exterior orientation parameters are not included in the metadata of the images.

New images can be added to the dataset simply by following the described structure of the annotation data shown in Fig. 4. This does not require any changes in the provided evaluation codes. New photo shoots, i.e., new settings of people, must be identified with a unique integer identifier. For any photo shoot, the real world locations of the people should stay the same in all the photos. There may be pictures taken from different camera locations. The person and camera tags should start with a letter P and C, respectively, followed by a unique identifier integer. The person and camera location tags must be consistent within a given photo shoot, however, repeated tags in different photo shoots are allowed. This means that two different people or camera tags can be the same as long as they belong to a different photo shoot. At least 1 of 4 body parts (center of the eyes, shoulders, torso, head) of the people in the images must be annotated in terms of pixel locations. They should be named "Eyes", "Shoulder", "Torso", and "Head" in the body part column of the body part pixel location file in Fig. 4(a).

To be consistent with the annotations in the provided test images, the annotation can be done as follows. Using the keypoint numbering in Fig. 6, the center of the eyes refers to the middle point of the keypoint pair 15–16, the center of the shoulders refers to the middle point of the keypoint pair 2–5, the center of the torso refers to the middle point of the keypoint pair 1–8, and the head should be annotated as middle point of the head regardless of the head's angle with respect to the camera. If a head is sideways and only one of the eyes is visible, the visible eye can be annotated as the center of the eyes. If no eyes are visible, the center of the eyes should not be annotated. The center of the eyes should also not be annotated if at least one of the eyes is out of the picture due to the head being on the edge of the picture. The other body parts can be annotated as long as they are either completely visible in the picture or are partially occluded by another person or object. In the cases where they are partially occluded, the pixel location should be estimated as if the occluding person or object was not present in the picture. The center of the shoulders, torso, and head should not be annotated only in the cases where these body parts are either partially or completely out of the picture due to the person being on the edge of the picture. If a person is sideways and only one of the shoulders, i.e., keypoints 2 and 5, is visible, this point can be annotated as the center of the shoulders.

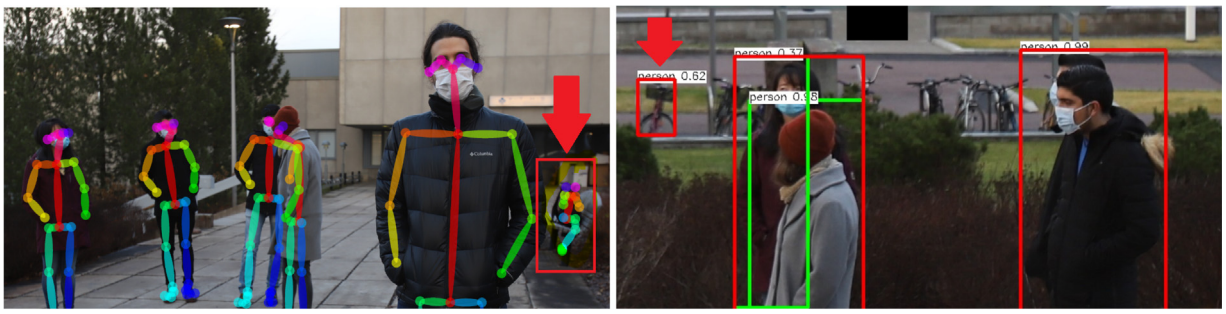

Fig. 5. False positive examples for OpenPose (left) and YOLOv4 (right). The images are from our test set.

3.3. Evaluation protocol

Any distance estimation method to be tested using the benchmark should give as output at least 1 of the 4 annotated pixel body locations along with either the estimated 3D location of the persons or the estimated distances between the people. The body part can be different for each person, or a method may choose to give only a single body part, such as the head, for all the persons. The test benchmark uses the pixel locations to automatically match each detected person with one of the ground truth locations and then computes average percentual pair-wise estimation errors between the estimated and ground truth distances.

We provide all the necessary functionalities for testing as long as the required output for each image is given. Internally, the matching is carried out by comparing the automatically detected body pixel locations with the points annotated in the files. The automatically detected body parts are compared to all of the respective annotated body parts. As an example, a detected torso point is compared to all of the annotated torso points for that image. For all of the detected body parts of a person, the closest respective annotated point in terms of pixel-wise distance is found. In case there are more than one detected persons matched with the same ground truth person, the matching is done in a greedy manner by selecting only the closest match and the rest of the detected persons for that ground truth person are regarded as false positives.

After matching the detections with the persons labeled in the photos, we calculate the distances between each person pair by using their estimated 3D locations. Then, the estimated pair-wise distances are compared to the corresponding ground truth pair-wise distances to obtain a percentual distance estimation error for each pair. The performance is evaluated by taking the average of all of the pair-wise percentual distance estimation errors for each image and then averaging over images. In addition to the pair-wise percentual distance estimation error, we evaluate also the person detection rate, i.e., the ratio of correctly detected person averaged over all the images, and the false discovery rate averaged over all the images. It should be noted here that we do not use any threshold for matching the detections with the actual people. As long as the number of detections is lower or equal to the actual number of people in an image, all the detections are matched. Thus, detections can be considered false positives only if there are more detections than actual people for an image. Therefore, a method producing many false positive detections is expected to get a high detection rate, but naturally the distance estimations would likely be poor and the false discovery rate would be higher. On the other hand, a method missing most the people could have a low pair-wise percentual distance estimation error for the detected people, but still not be suitable for social distancing analysis. Therefore, it is important to consider all these metrics together, when evaluating a social distance estimation algorithm.

The pair-wise percentual distance estimation error D_e for the eth single image is given by the following formula, where n is the number of detected people in the image, E_i is the estimated 3D location of the ith person and G_i is the ground truth 3D location of the ith person:

$$D_e = \frac{\sum_{k=1}^{n-1}\sum_{i=k+1}^{n} \frac{\left|\|E_k - E_i\| - \|G_k - G_i\|\right|}{\|G_k - G_i\|} * 100}{\binom{n}{2}}. \quad (1)$$

Here, the distances may be also directly given instead of the 3D locations.

In order to obtain an overall distance estimation error metric for a set of images, D_e of all of the images in the image set are averaged. The distance estimation error for a set of images D_E is given by the following formula where N is the number of images in the set:

$$D_E = \frac{\sum_{e=1}^{N} D_e}{N}. \quad (2)$$

The test benchmark gives D_E, the person detection rate, and the false discovery rate as an output for a given set of images as long as the input and annotated data are provided in the proper format. Currently, the test benchmark uses our provided test photos, but if new images are added to the dataset as explained in Section 3.2, these will be automatically considered in the evaluation.

4. Proposed method for social distance estimation

Our proposed method to estimate social distances takes advantage of deep learning-based object detection and human pose estimation methods. Firstly, the input image is given to YOLOv4 (Bochkovskiy et al., 2020) object detection model to obtain bounding boxes for people. After bounding boxes are obtained, overlapping boxes are grouped together. Then, these grouped boxes are cropped from the full image and they are individually given to OpenPose (Cao et al., 2019) human pose estimation model. After the skeleton keypoints are extracted from OpenPose, the pixel locations of these keypoints are used in our distance estimation algorithm to obtain 3D location estimates for each person in the image.

When YOLOv4 and OpenPose models are used together, they eliminate each other's false positives. The left image in Fig. 5 shows a case where a backpack is falsely recognized as a human by OpenPose. However, YOLOv4 does not recognize it as a human. Therefore, the backpack would not be cropped and given to the OpenPose model. The right image in Fig. 5 shows a case where a bicycle is falsely recognized as a human by the YOLOv4 model. The bicycle is then cropped from the full image and given to the OpenPose model. However, the OpenPose model does not detect any human skeleton in the cropped bicycle image. Therefore, neither of these false positive cases is further processed by the distance estimation algorithm.

After the cropped images from YOLOv4 are processed by the OpenPose model, the skeleton keypoints for detected human bodies are extracted. We use the 25 keypoint output version of OpenPose illustrated in Fig. 6. Out of the extracted keypoints, we select pairs whose mutual distance is independent of the person's pose, whose average distance is available in the literature, whose angle towards the lens is

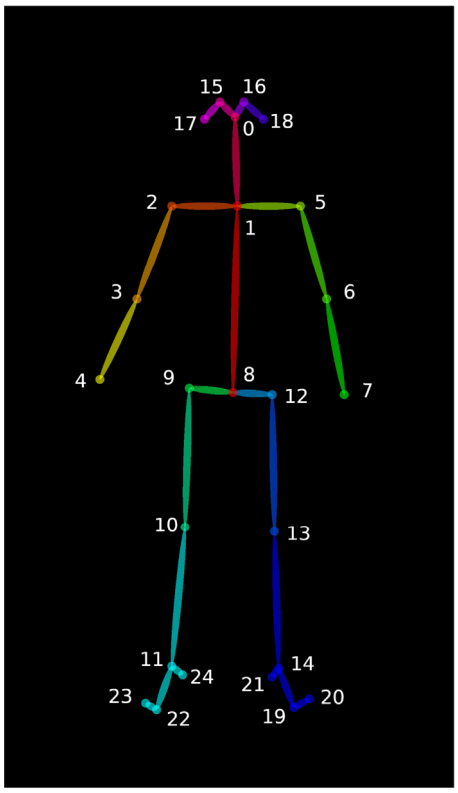

Fig. 6. 25 skeleton keypoint output of OpenPose.

as constant as possible, and which are visible in most of the photos. With these criteria, we select three key point pairs for our algorithm: 15–16 for pupillary distance, 2–5 for shoulder width, and 1–8 for torso length. In typical media or personal photos, the torso has the most constant angle towards the lens, but the eyes and shoulders are visible also in the close-up and portrait photos, where the torso is not seen. We assume average adult body proportions for the three keypoint pairs: 389 mm for shoulder width (Watson, 2018), 63 mm for pupillary distance (Evans, 2019), and 444 mm for torso length (White Mountain Backpacks, 2021). The extracted keypoint pairs are then processed by our distance estimation algorithm that estimates 3D positions with respect to the camera for each person.

We use the pinhole camera model (Sturm, 2014) shown in Fig. 7 for our calculations. We also make an assumption that every keypoint pair is parallel to the camera's sensor plane. We make these assumptions because the subjects' poses and camera's exterior orientation parameters (Zhang, 2014) are not known. Estimating the exterior orientation parameters (Zhang, 2014) of the camera from single images is an ill-posed problem (Kabanikhin, Tikhonov, Ivanov, & Lavrentiev, 2008), but in most cases the angle between a person's torso and the camera's sensor plane is negligible for our calculations.

We denote 3D locations of the keypoints on the image coordinate system as

$$(x_a, y_a, f), \tag{3}$$

where f is the focal length, and 3D location estimates of the keypoints on the world coordinate system as

$$E_n = (X_a, Y_a, -d), \tag{4}$$

where d is the distance to the camera. The distance between a pair of keypoints on the image coordinate system is

$$D_i = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (f - f)^2} \tag{5}$$

and the distance between the keypoints on the world coordinate system is

$$D_w = \sqrt{(X_0 - X_1)^2 + (Y_0 - Y_1)^2 + (d - d)^2}. \tag{6}$$

Since the camera sensor's plane size is known, x_a and y_a in Eq. (3) can be derived from the x and y pixel locations of the keypoints in the image. The last coordinate, f, in Eq. (3) is obtained from the camera parameters. Thus, all the keypoints' 3D positions on the image coordinate system in Eq. (3) are known and D_i can be solved. By using triangle similarity, the following equations give 3D positions of the keypoints on the world coordinate system. Eq. (7), where D_w is one of the average body proportions, is used to derive d in Eq. (4). After d is derived, X_a and Y_a are obtained from Eqs. (8) and (9):

$$\frac{D_i}{f} = \frac{D_w}{d} \tag{7}$$

$$X_a = -\frac{d}{f} x_a \tag{8}$$

$$Y_a = -\frac{d}{f} y_a \tag{9}$$

After the 3D coordinates of the keypoints on the world coordinate system in Eq. (4) are estimated, the middle points of each detected keypoint pair are used to represent a 3D location for the person. Thus, we have at most 3 different estimated 3D locations for a person, one for each keypoint pair (shoulder, pupil, torso). While we assume that the keypoint pairs are parallel to the camera's sensor plane, this assumption may not be valid, and the accuracy of the estimated locations is affected by the severity of the violations. Fig. 8 shows the birdseye view of a person's orientation angle θ towards the lens. If the angle is non-

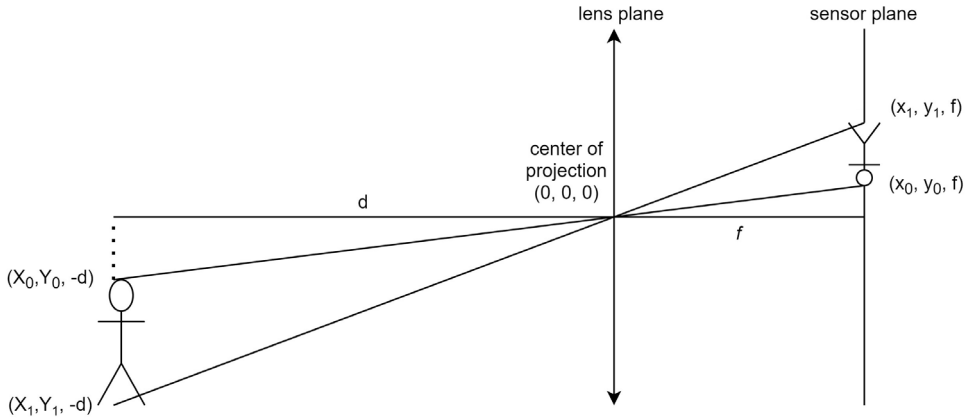

Fig. 7. Pinhole camera model.

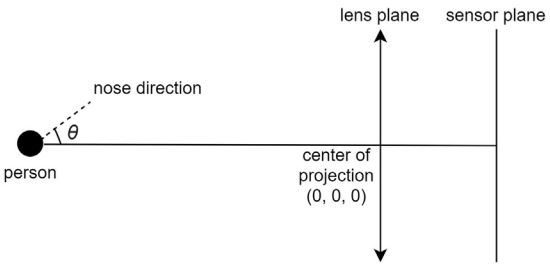

Fig. 8. Birdseye view of orientation angle towards the lens.

zero, the shoulder and pupil keypoint pairs are no longer parallel to the sensor plane and the estimates based on these keypoint pairs are prone to error. However, in a typical situation of upright torsos the estimates made from the torso length are unaffected by θ, because θ does not affect D_i computed using Eq. (5) for the torso. On the other hand, also a torso may not be parallel to the sensor plane either because the person is in a bent position or because the camera's pitch angle is non-zero. For an overhead image, shoulders might be parallel to the sensor plane, while torsos would be perpendicular. Whenever the assumption on a keypoint pair being parallel to the sensor plane is violated, D_i in Eq. (5) decreases. A smaller D_i leads to a larger estimate for d from Eq. (7). For this reason, we select the 3D location estimate with the smallest distance to the camera. For typical media or personal photos, where the pitch angle is small, this usually means using the estimate derived from the torso whenever it is available. However, for close-up and portrait pictures, the torso is often not visible. Fig. 9 shows three pictures taken from the same location but with increasing focal lengths. The rightmost image in Fig. 9 is an example of a close-up picture where the distance estimations have to be made from the shoulder and pupil distances since there are no visible torsos.

Finally, our method computes the distances between all the pairs of detected people and gives them as outputs. The pixel locations for the detected persons are given to be able to evaluate on our benchmark, while they are not needed if the method is used for analyzing social distancing in novel images for photographic studies. The overall flowchart of the proposed social distance estimation method is illustrated in Fig. 10.

5. Experimental results

5.1. Experimental setup

All of the code was developed in Python programming language version 3.8 (Van Rossum & Drake Jr, 1995). OpenPose (Cao et al., 2019) and YOLOv4 (Bochkovskiy et al., 2020) models were used for human detection and pose estimation. The input size of YOLOv4 was set to 704 × 704. Input size was not set for OpenPose as OpenPose is able to adapt its input size for each image. The version of the OpenPose model we were using was originally trained by using the COCO keypoint challenge dataset (Lin et al., 2014), combined with OpenPose authors' own annotated dataset for foot keypoint estimation which consists of a small subset of the COCO dataset where the authors labeled foot keypoints. YOLOv4 uses CSPDarknet53 (Wang et al., 2019) as its backbone which was trained on the ImageNet dataset (Deng et al., 2009). The deep learning models were downloaded from their respective official source code pages[2,3] and they were loaded and used by TensorFlow library version 2.3.1 (Abadi et al., 2015). For image processing purposes, OpenCV imaging library was used (Bradski, 2000). In addition to our final method that generates 3D position estimates using torso, shoulders, and eyes and selects the estimate closest to the camera as explained in Section 4, we also evaluate variants of the proposed method, where only one of these body parts is used at the time. We use our test benchmark to compute the results for all the images and for each photo shoot separately.

5.2. Results

Table 2 shows the person detection rates and pair-wise percentual distance estimation errors for the overall dataset. Table 3 gives the results for the first photo shoot separately. For the other photo shoots, the separate results are provided in the Appendix B.6–B.8. Since YOLOv4 is used in addition to OpenPose and they cancel each other's false positives, we have no cases with more detections than actual people in an image. This leads to almost zero false discovery rates as explained in Section 3.3. Therefore, false discovery rates are not reported in the tables.

It can be observed from Table 2 that the most reliable body part to estimate locations is the torso. However, estimations made from the torso alone fail for close-up pictures where the torso detection rate is low. When all three body parts (shoulder, pupil, and torso) are used together for the estimations, the obtained results shown in the last column are better than the results obtained from any single body part. The combined method mostly uses the torso whenever it is visible (overall shots) and uses the shoulder and pupil distances when the torso is not visible (close-up shots).

Looking at Tables 3, B.6, B.7 and B.8 it can be seen that there are no significant differences in terms of person detection rates when it comes to indoor and outdoor pictures. However, it should be noted that the pair-wise distance estimation errors for the indoor pictures are slightly higher than the outdoor pictures. This is primarily caused by the fact that many body parts of the people in the indoor pictures are obstructed by the chairs and sofas. There are also more cases of people facing away from the camera, people standing in front of other people, and people in poses where their torsos were non-upright in the indoor photo shoots.

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose
[3] https://github.com/AlexeyAB/darknet

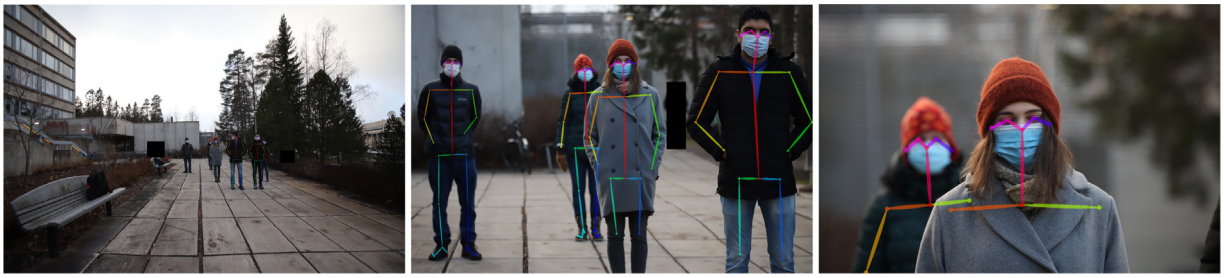

Fig. 9. Examples of pictures from the dataset belonging to the first photo shoot, all of them taken from camera location C1. The used focal lengths for the pictures are 16 mm, 105 mm and 300 mm from left to right.

Table 2
Person detection rates and pair-wise percentual distance errors for each of the methods for all of the images (indoor and outdoor) combined.

Focal length (mm)	Number of pictures	Shoulder based method		Pupil based method		Torso based method		Combined method	
		Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error
16	28	0.75	333.42	0.55	39.79	0.82	36.30	**0.89**	**28.80**
24	46	0.81	346.05	0.55	39.52	0.91	33.22	**0.94**	**24.68**
35	48	0.81	450.49	0.58	65.63	0.91	48.52	**0.92**	**34.68**
50	84	0.80	306.56	0.44	72.37	0.91	39.29	**0.94**	**35.03**
105	69	0.72	332.72	0.57	110.50	0.79	73.29	**0.89**	**52.50**
200	7	0.69	105.28	0.73	52.28	0.69	93.53	**0.78**	**53.66**
300	18	0.70	1244.59	0.60	52.88	0.61	148.94	**0.78**	**52.51**
All	300	0.78	385.22	0.54	68.56	0.84	51.01	**0.91**	**38.24**

Table 3
Person detection rates and pair-wise percentual distance errors for each of the methods for the first photo shoot (outdoor) where every person is standing up.

Focal length (mm)	Number of pictures	Shoulder based method		Pupil based method		Torso based method		Combined method	
		Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error
16	7	**0.85**	120.60	0.71	26.44	**0.85**	**18.33**	**0.85**	18.48
24	8	0.83	190.70	0.64	76.24	**0.91**	**16.99**	**0.91**	21.49
35	11	0.90	174.68	0.84	57.78	**0.96**	**20.17**	**0.96**	21.09
50	11	0.87	190.12	0.77	72.35	0.89	**24.34**	**0.91**	26.40
105	11	**1.00**	127.57	**1.00**	48.99	**1.00**	41.63	**1.00**	**33.08**
200	7	0.69	105.28	0.73	**52.28**	0.69	93.53	**0.78**	53.66
300	8	0.70	288.13	0.88	**34.48**	0.18	–	**0.89**	**34.48**
All	63	0.85	165.27	0.78	54.43	0.90	**28.76**	**0.91**	28.97

Table 4
Person detection rates and pair-wise percentual distance errors for the combined method for the photos taken from camera location C2, for which the zero pitch angle assumption is not valid.

Number of pictures	Combined method	
	Person detection rate	Pair-wise percent distance error
53	0.85	37.59

Table 5
F1-scores of our proposed method for different safe distance thresholds.

Safe distance (m)	F1-score
1	0.46
1.5	0.62
2	0.75
3	0.83
4	0.90

5.3. Additional results and analysis

We separately show the results for the images that were taken from camera location C2 for the first and third photo shoot (outdoor) on Table 4. C2 location was at a height of 360 cm on the first and 220 cm on the third photo shoot relative to the ground plane where the subjects were standing on. Thus, the camera was pitched down to include the subjects within the field of view. For the other camera locations, the pitch angle was close to zero and people were mainly standing or sitting with their torsos upright. Therefore, the torsos are usually almost parallel to the camera's sensor plane and, thus, produce good distance estimates whenever they are visible. For camera location C2, this may no longer be the case. However, the results show that the relative pair-wise distance estimation errors for C2 locations are slightly lower than on the average despite the violation of the zero pitch angle assumption. We can conclude that this level of pitch angle does not cause significant problems.

We also take a closer look on how errors vary for different pairs of people. Fig. 11 shows all the percent distance errors as a function of the corresponding ground-truth pair wise distance, i.e., each column of points corresponds to the different estimations for a specific pair of people. The variations for a specific pair may follow from different factors, such as the camera distance and angle, focal length, occlusions, and pose differences. It can be observed from this figure that the pair-wise distance estimations errors and error variations are on average lower for higher ground truth distances. This is reasonable as for the closest distances smaller absolute errors lead to higher percent errors and, therefore, the variations in the poses can also cause considerable percent error.

Furthermore, we also provide additional results by formulating the social distance estimation problem as a binary classification task similar to previous works. We set five different social distance thresholds as safe distances. If the distance between a pair is smaller than the threshold, we consider the distance to be unsafe and safe otherwise. We consider the unsafe case as the positive class. The standard evaluation metrics for binary classification problems are Precision, Recall, and F1-Score. The formulas for these metrics are $Precision = \frac{TruePositives}{TruePositives+FalsePositives}$, $Recall = \frac{TruePositives}{TruePositives+FalseNegatives}$, $F1-score = 2 * (\frac{Precision*Recall}{Precision+Recall})$. F1-score is an overall measure of the binary classification performance and is always within the range of 0–1 with 1 indicating perfect performance. The F1-score results of our proposed method are given in Table 5.

As can be seen in Table 5, the choice of safe distance threshold changes the F1-scores drastically. For example, the low performance for 1 m threshold follows from many ground-truth distances being just slightly above the threshold. As our methods tends to slightly underestimate the distances especially when the torsos are not visible as explained in Section 4, these cases lead to false positives. This supports our claim that formulating the problem of social distance estimation as a binary classification task is not an optimal way to evaluate the performance of the methods. As the results depend greatly on the threshold value, F1-scores do not reflect the true capacity and accuracy of the distance estimation performance of a method. Our proposed evaluation protocol, which gives the average pair-wise percentual distance estimation error offers greater insight on the method's performance.

6. Conclusion

To address the need for more accurate estimation of social distances from general images to analyze social and cultural impacts of the social

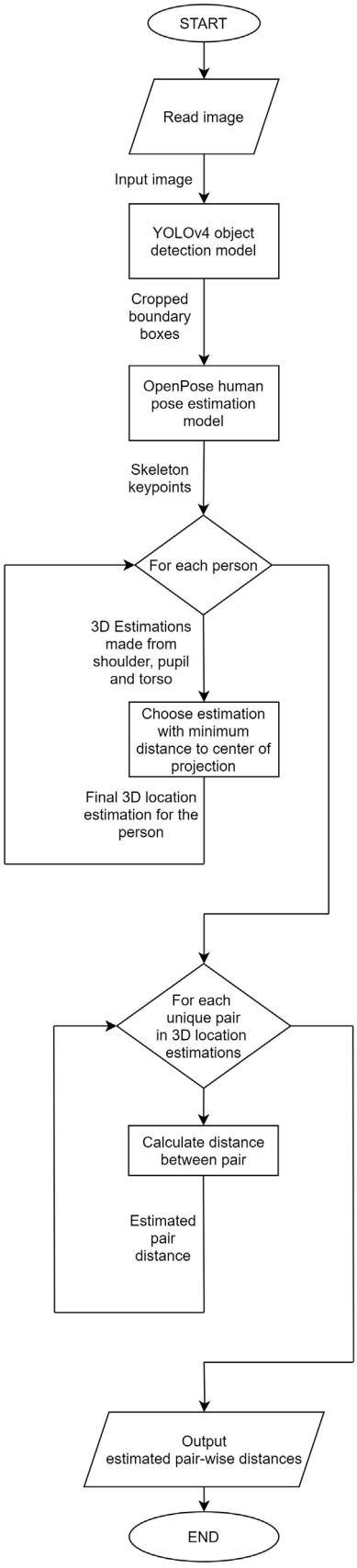

Fig. 10. Flowchart of the method.

distancing regulations introduced due to the COVID-19 pandemic, we proposed a new test benchmark for automatic social distance estimation algorithms. The benchmark includes an evaluation protocol for methods producing pair-wise social distances. The images follow a typical journalistic photographing style instead of a fixed monitoring setup, and they were taken with varying camera settings. Furthermore, we proposed a robust method that estimates 3D locations of persons in images and then uses these estimated locations to calculate the social distances between the people. Our method is able to estimate social distances in any single image without the need for knowing the extrinsic parameters or manually calibrating the homography matrix of the image plane to the ground plane, provided that the focal length and sensor size information of the camera are known, which enables our method to be used flexibly on all kinds of images. The proposed method was able to obtain 91% person detection rate along with 38.24% pair-wise distance error on the proposed test benchmark.

Main limitations of our proposed method follow from the assumptions made: at least one of keypoint pairs (eyes, shoulders, torso) is assumed to be parallel to the camera's sensor plane and the keypoint distances are assumed to follow average adult human body proportions. If these assumptions are violated, the accuracy of the proposed method will be affected. If the keypoint pairs are not parallel to the camera's sensor plane, this leads to distance estimates that are longer than the ground-truth. While in typical journalistic photos the camera's pitch angle is close to zero and the peoples' torsos are in upright positions, the torsos are not always visible. In particular, in close-up shots it is often necessary to make the estimations using either eye or shoulder keypoint pair, which are more commonly not following the parallelity assumption. Indeed, our experimental results showed satisfactory results for overall shots where the torsos of the people can be detected by OpenPose, but the accuracy of the estimations got weaker for close-up shots where the torsos were not visible in the image. Thus, our method could be further improved by estimating automatically also the pitch angle and persons' angles with respect to the camera. Due to the use of average adult human body proportions, the estimations made for children in the images would be less accurate. This problem could be tackled by taking advantage of other methods that can estimate the gender and ages of the subjects and adaptively changing the assumed body dimensions for each individual subject depending on their gender and age. Furthermore, our method requires the focal length and sensor plane size information of the camera and cannot be applied on photos where these information are lacking. For our method to be applied on images where the focal length and sensor plane size are not known, they would have to be estimated through other methods.

It should be remembered that our approach is not intended for online monitoring of social distances. For such purposes, there are multiple approaches proposed in the literature taking advantage of additional sensors and/or fixed monitoring setup. Instead, our work was motivated by the need to analyze long term changes in average social distances caused by COVID-19 pandemic using personal or media photo collections, where the imaging setup can vary significantly and no additional sensor information can be obtained. This kind of analysis will require comparing tens of thousands pre-pandemic and post-pandemic photos to draw any statistically significant conclusions, while individual images and distances are not relevant. Furthermore, it is not meaningful to define exact average social distances, but rather look into approximate percentual change. Thus, if we can assume that similar errors occur for both pre-pandemic and post-pandemic photos, potentially interesting conclusions can be made already with the current accuracy of the approach despite the above-mentioned limitations.

In our future research, we will use our benchmark to further enhance the proposed method and then use it in an interdisciplinary study, where we will analyze the impacts of the COVID-19 regulations on social interactions. To this end, it will be important to verify that the pre-pandemic and post-pandemic are large enough and similar enough

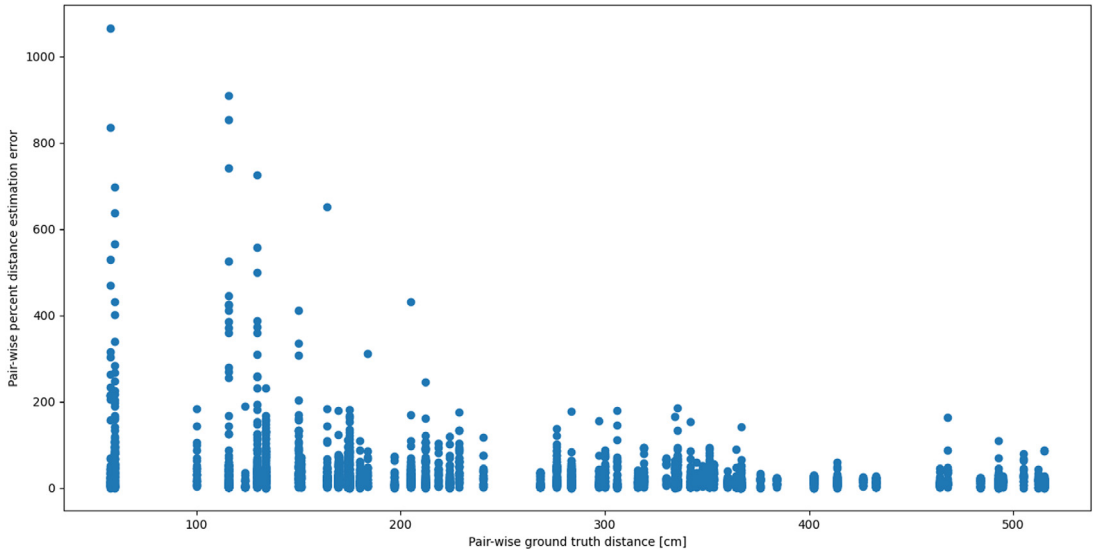

Fig. 11. Pair-wise distance estimation errors for each of the ground truth pair-wise distances.

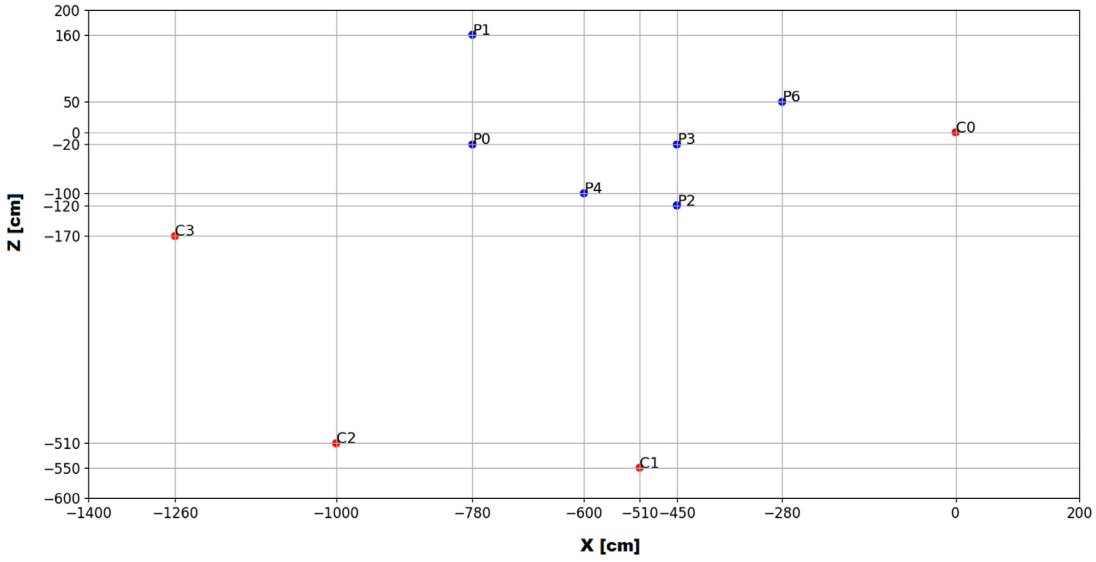

Fig. A.12. Birdseye view of the second photo shoot (indoor). The ground truth locations of the people and cameras are given in blue and red dots, respectively.

so that errors can be assumed to occur at similar rates and some statistically significant conclusions can be drawn. While the COVID-19 makes the social distance analysis very topical, the benchmark and the developed methods are naturally not restricted on COVID-19 related analysis, but they can be beneficial in other image-based proxemics studies focusing on different historical, cultural, or journalistic phenomena.

CRediT authorship contribution statement

Mert Seker: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft. **Anssi Männistö:** Conceptualization, Investigation, Resources, Writing – review & editing, Supervision, Project Administration, Funding acquisition. **Alexandros Iosifidis:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Jenni Raitoharju:** Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – review & editing, Supervision, Project Administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Our dataset and all the evaluation codes are publicly available at https://doi.org/10.23729/b2ea87e6-b845-46b8-abf3-cdbe299ce8b0 (requires signing a license agreement). All the evaluation codes are available without any the license agreement also at https://github.com/mertseker-dev/social-distance-estimation.

Appendix A. Birdseye views of photo shoots 2–4

See Figs. A.12–A.14.

Appendix B. Results for photo shoots 2-4

See Tables B.6–B.8.

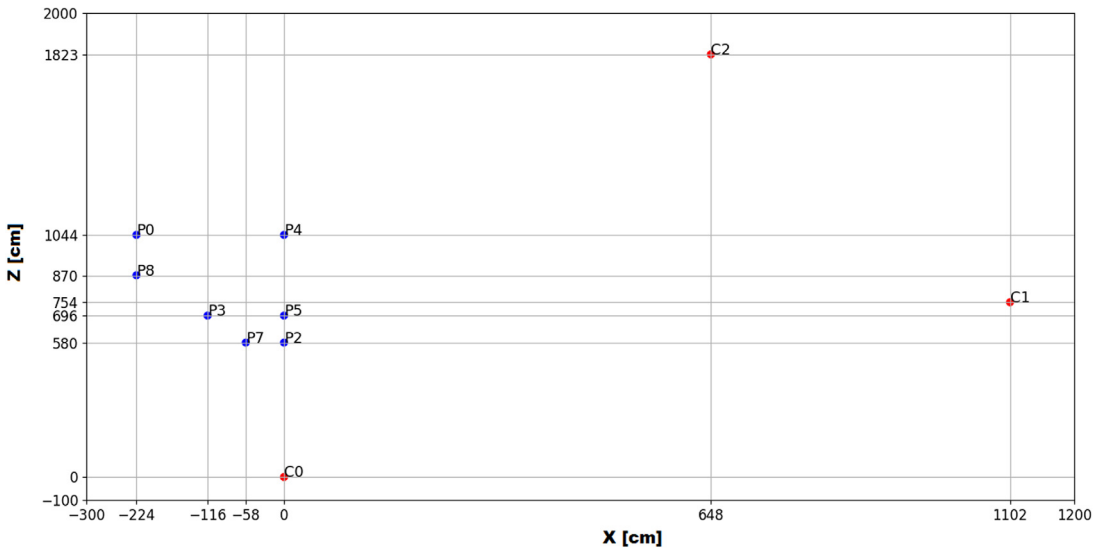

Fig. A.13. Birdseye view of the third photo shoot (outdoor). The ground truth locations of the people and cameras are given in blue and red dots, respectively.

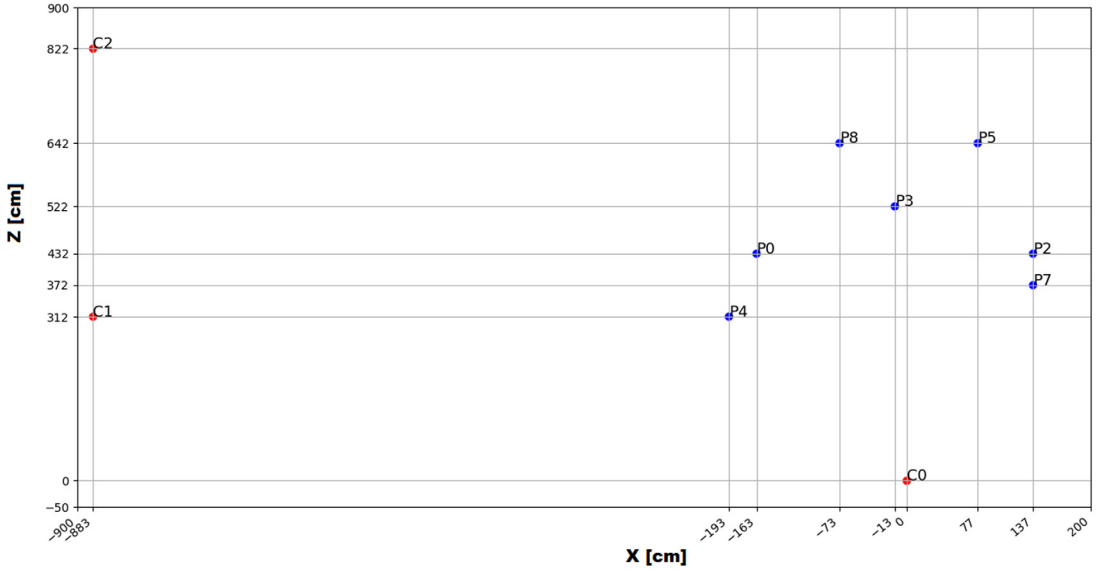

Fig. A.14. Birdseye view of the fourth photo shoot (indoor). The ground truth locations of the people and cameras are given in blue and red dots, respectively.

Table B.6
Person detection rates and pair-wise percentual distance errors for each of the methods for the second photo shoot (indoor) where every person is sitting down.

Focal length (mm)	Number of pictures	Shoulder based method		Pupil based method		Torso based method		Combined method	
		Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error
16	4	0.83	178.58	0.62	29.70	0.62	**22.92**	**1.00**	23.98
24	4	0.83	354.18	0.54	24.94	0.70	**19.29**	**0.95**	23.40
35	4	0.66	49.27	0.54	**19.61**	**0.95**	25.79	**0.95**	20.40
50	7	0.76	189.61	0.51	29.79	0.76	29.57	**0.89**	**27.26**
105	14	0.68	102.84	0.62	55.40	0.56	**27.42**	**0.90**	35.07
All	33	0.74	163.61	0.57	37.76	0.70	**26.03**	**0.93**	28.88

Table B.7
Person detection rates and pair-wise percentual distance errors for each of the methods for the third photo shoot (outdoor) where every person is standing up.

Focal length (mm)	Number of pictures	Shoulder based method		Pupil based method		Torso based method		Combined method	
		Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error
16	9	0.62	760.07	0.42	51.13	0.81	**18.66**	**0.84**	28.78
24	8	0.84	778.06	0.33	33.50	0.91	**17.06**	**0.93**	22.37
35	13	0.78	880.81	0.40	68.43	0.84	**16.50**	**0.86**	19.09
50	50	0.83	333.23	0.33	82.12	0.96	**25.18**	**0.97**	32.54
105	21	0.70	771.80	0.37	149.17	0.84	**36.14**	**0.88**	67.45
300	10	0.70	1669.68	0.41	117.25	0.63	148.94	**0.73**	**66.52**
All	111	0.78	658.29	0.34	81.44	0.88	**34.21**	**0.91**	39.35

Table B.8
Person detection rates and pair-wise percentual distance errors for each of the methods for the fourth photo shoot (indoor) where some people are sitting down and some are standing up.

Focal length (mm)	Number of pictures	Shoulder based method		Pupil based method		Torso based method		Combined method	
		Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error	Person detection rate	Pair-wise percent distance error
16	8	0.78	117.07	0.54	42.86	0.9	82.69	**0.92**	**40.27**
24	26	0.79	259.67	0.60	31.07	0.95	45.33	**0.96**	**26.57**
35	20	0.82	402.73	0.61	76.93	**0.94**	89.47	**0.94**	**55.15**
50	16	0.74	429.24	0.52	74.61	0.8	98.24	**0.86**	**54.40**
105	23	0.68	102.05	0.57	155.77	0.80	137.55	**0.85**	**57.06**
All	93	0.76	266.61	0.58	79.10	0.89	90.03	**0.91**	**46.22**

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. URL: http://tensorflow.org/.

Abouk, R., & Heydari, B. (2021). The immediate effect of COVID-19 policies on social-distancing behavior in the United States. *Public Health Reports*, *136*(2), 245–252. http://dx.doi.org/10.1177/0033354920976575.

Aghaei, M., Bustreo, M., Wang, Y., Bailo, G. L., Morerio, P., & Del Bue, A. (2021). Single image human proxemics estimation for visual social distancing. In *IEEE winter conference on applications of computer vision* (pp. 2785–2795). http://dx.doi.org/10.1109/WACV48630.2021.00283.

Ahmed, I., Ahmad, M., Rodrigues, J. J., Jeon, G., & Din, S. (2021). A deep learning-based social distance monitoring framework for COVID-19. *Sustainable Cities and Society*, *65*, Article 102571. http://dx.doi.org/10.1016/j.scs.2020.102571.

Al-Hasan, A., Khuntia, J., & Yim, D. (2020). Threat, coping, and social distance adherence during COVID-19: Cross-continental comparison using an online cross-sectional survey. *Journal of Medical Internet Research*, *22*(11), http://dx.doi.org/10.2196/23019.

Al-Hasan, A., Yim, D., & Khuntia, J. (2020). Citizens' adherence to COVID-19 mitigation recommendations by the government: A 3-country comparative evaluation using web-based cross-sectional survey data. *Journal of Medical Internet Research*, *22*(8), http://dx.doi.org/10.2196/20634.

Balasa, A. P. (2020). COVID – 19 on lockdown, social distancing and flattening the curve – A review. *European Journal of Business and Management Research*, *5*(3), http://dx.doi.org/10.24018/ejbmr.2020.5.3.316.

Benfold, B., & Reid, I. (2011). Stable multi-target tracking in real-time surveillance video. In *IEEE conference on computer vision and pattern recognition* (pp. 3457–3464). http://dx.doi.org/10.1109/CVPR.2011.5995667.

Bertoni, L., Kreiss, S., & Alahi, A. (2021). Perceiving humans: From monocular 3D localization to social distancing. *IEEE Transactions on Intelligent Transportation Systems*, http://dx.doi.org/10.1109/TITS.2021.3069376.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, URL: https://opencv.org/.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., & Sheikh, Y. A. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, http://dx.doi.org/10.1109/TPAMI.2019.2929257.

Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., et al. (2018). WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection. In *IEEE conference on computer vision and pattern recognition*. http://dx.doi.org/10.1109/CVPR.2018.00528.

Cook, M. (1970). Experiments on orientation and proxemics. *Human Relations*, *23*(1), 61–76. http://dx.doi.org/10.1177/001872677002300107.

Courtemanche, C., Garuccio, J., Le, A., Pinkston, J., & Yelowitz, A. (2020). Strong social distancing measures in the United States reduced the COVID-19 growth rate. *Health Affairs*, *39*(7), 1237–1246. http://dx.doi.org/10.1377/hlthaff.2020.00608.

Cristani, M., Bue, A. D., Murino, V., Setti, F., & Vinciarelli, A. (2020). The visual social distancing problem. *IEEE Access*, *8*, 126876–126886. http://dx.doi.org/10.1109/ACCESS.2020.3008370.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: a large-scale hierarchical image database. (pp. 248–255). http://dx.doi.org/10.1109/CVPR.2009.5206848,

Di Corrado, D., Magnano, P., Muzii, B., Coco, M., Guarnera, M., De Lucia, S., et al. (2020). Effects of social distancing on psychological state and physical activity routines during the COVID-19 pandemic. *Sport Sciences for Health*, *16*(4), 619–624. http://dx.doi.org/10.1007/s11332-020-00697-5.

Doogan, C., Buntine, W., Linger, H., & Brunt, S. (2020). Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across Six Countries: A topic modeling analysis of Twitter data. *Journal of Medical Internet Research*, *22*(9), http://dx.doi.org/10.2196/21419.

Eden, A. L., Johnson, B. K., Reinecke, L., & Grady, S. M. (2020). Media for coping during COVID-19 social distancing: Stress, anxiety, and psychological well-being. *Frontiers in Psychology*, *11*, 3388. http://dx.doi.org/10.3389/fpsyg.2020.577639.

Evans, L. (2019). What is pupillary distance and how do you measure it?. Online: https://www.allaboutvision.com/eye-care/measure-pupillary-distance/ Accessed: 03- Mar- 2021.

Fabbri, M., Lanzi, F., Gasparini, R., Calderara, S., Baraldi, L., & Cucchiara, R. (2020). Inter-homines: Distance-based risk estimation for human safety. arXiv:2007.10243.

Fleuret, F., Berclaz, J., Lengagne, R., & Fua, P. (2008). Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 267–282. http://dx.doi.org/10.1109/TPAMI.2007.1174.

Ford, M. B. (2020). Social distancing during the COVID-19 pandemic as a predictor of daily psychological, social, and health-related outcomes. *The Journal of General Psychology*, 1–23. http://dx.doi.org/10.1080/00221309.2020.1860890.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(59), 1–35.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE conference on computer vision and pattern recognition* (pp. 3354–3361). http://dx.doi.org/10.1109/CVPR.2012.6248074.

Hall, E. T. (1966). *The hidden dimension* (1st ed.). Doubleday Garden City, N.Y.

Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannan, P., Diebold, A. R., Durbin, M., et al. (1968). Proxemics [and comments and replies]. *Current Anthropology*, *9*(2/3), 83–108. http://dx.doi.org/10.1086/200975.

Harrigan, J. A. (2005). Proxemics, kinesics, and gaze. In *The new handbook of methods in nonverbal behavior research*. Oxford University Press.

Jacob, L., Tully, M. A., Barnett, Y., Lopez-Sanchez, G. F., Butler, L., Schuch, F., et al. (2020). The relationship between physical activity and mental health in a sample of the UK public: A cross-sectional study during the implementation of COVID-19 social distancing measures. *Mental Health and Physical Activity*, *19*, Article 100345. http://dx.doi.org/10.1016/j.mhpa.2020.100345.

Kabanikhin, S., Tikhonov, N., Ivanov, V., & Lavrentiev, M. (2008). Definitions and examples of inverse and ill-posed problems. *Journal of Inverse and Ill-Posed Problems*, *16*, 317–357.

Lee, M., Kang, B., & You, M. (2021). Knowledge, attitudes, and practices (KAP) toward COVID-19: a cross-sectional study in South Korea. *BMC Public Health*, *21*(295), http://dx.doi.org/10.1186/s12889-021-10285-y.

Lin, T.-Y., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., et al. (2014). Microsoft COCO: common objects in context. arXiv:1405.0312.

Morerio, P., Bustreo, M., Wang, Y., & Bue, A. D. (2021). End-to-end pairwise human proxemics from uncalibrated single images. In *IEEE international conference on image processing* (pp. 3058–3062). http://dx.doi.org/10.1109/ICIP42928.2021.9506457.

Nguyen, C. T., Saputra, Y. M., Van Huynh, N., Nguyen, N.-T., Khoa, T. V., Tuan, B. M., et al. (2020). A comprehensive survey of enabling and emerging technologies for social distancing—Part II: Emerging technologies and open issues. *IEEE Access*, *8*, 154209–154236. http://dx.doi.org/10.1109/access.2020.3018124.

Pouw, C. A., Toschi, F., van Schadewijk, F., & Corbetta, A. (2020). Monitoring physical distancing for crowd management: Real-time trajectory and group analysis. *PLoS One*, *15*(10), Article e0240963. http://dx.doi.org/10.1371/journal.pone.0240963.

Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., et al. (2020). The effect of control strategies that reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.3552864.

Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2020). Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques. arXiv:2005.01385.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv:1804.02767.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv:1506.01497.

Sturm, P. (2014). Pinhole camera model. In *Computer vision, a reference guide*. Springer US, http://dx.doi.org/10.1007/978-0-387-31439-6_472.

Sun, C., & Zhai, Z. (2020). The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission. *Sustainable Cities and Society*, *62*, Article 102390. http://dx.doi.org/10.1016/j.scs.2020.102390.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial*. The Netherlands: Centrum voor Wiskunde en Informatica Amsterdam.

Vokó, Z., & Pitter, J. G. (2020). The effect of social distance measures on COVID-19 epidemics in Europe: an interrupted time series analysis. *GeroScience*, *42*(4), 1075–1082. http://dx.doi.org/10.1007/s11357-020-00205-0.

Wang, C.-Y., Liao, H.-Y. M., Yeh, I.-H., Wu, Y.-H., Chen, P.-Y., & Hsieh, J.-W. (2019). CSPNet: A new backbone that can enhance learning capability of CNN. arXiv:1911.11929.

Watson, K. (2018). What's an average shoulder width?. Online: https://www.healthline.com/health/average-shoulder-width Accessed: 03- Mar- 2021.

White Mountain Backpacks (2021). Backpack Fitting. Online: https://www.whitemountain.com.au/backpack-fitting/backpack-fitting-measure-torso-length.html Accessed: 03- Mar- 2021.

Wojke, N., & Bewley, A. (2018). Deep cosine metric learning for person re-identification. In *IEEE winter conference on applications of computer vision* (pp. 748–756). http://dx.doi.org/10.1109/WACV.2018.00087.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *IEEE international conference on image processing* (pp. 3645–3649). http://dx.doi.org/10.1109/ICIP.2017.8296962.

Yang, D., Yurtsever, E., Renganathan, V., Redmill, K. A., & Özgüner, Ü. (2020). A vision-based social distancing and critical density detection system for COVID-19. arXiv:2007.03578.

Young, J. W. (1982). *Projective geometry*. Pub. for the Mathematical Association of America by the Open Court Pub. Co..

Zhang, Z. (2014). Camera parameters (internal, external). In *Computer vision: A reference guide* (p. 81). Boston, MA: Springer US, http://dx.doi.org/10.1007/978-0-387-31439-6_100019.

Zhang, W., Gao, F., Gross, J., Shrum, L. J., & Hayne, H. (2021). How does social distancing during COVID-19 affect negative moods and memory? *Memory*, *29*(1), 90–97. http://dx.doi.org/10.1080/09658211.2020.1857774.

Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. arXiv:1904.07850.