Single View Physical Distance Estimation using Human Pose Supplementary Material

Xiaohan Fei, Henry Wang, Lin Lee Cheong, Xiangyu Zeng*, Meng Wang, and Joseph Tighe
Amazon Web Services

August 17, 2021

Contents

1 Derivation

1.1 Linear Constraint on W

In this section, we derive the *linear scale-invariant constraint* on $\mathbf{W}$ in Eq (3) of the main paper. We repeat the projection equation of the ankle center point of person i and j here for completeness:

$$\begin{aligned} \lambda_{B,i}\bar{\mathbf{x}}_{B,i} &= \mathbf{K}\mathbf{X}_{B,i} \\ \lambda_{B,j}\bar{\mathbf{x}}_{B,j} &= \mathbf{K}\mathbf{X}_{B,j}. \end{aligned} \tag{1}$$

*Work was done when Xiangyu was at Amazon.

Take the difference of the two equations above, and multiply both sides by $\mathbf{N}^\top\mathbf{K}^{-1}$, we have

$$\mathbf{N}^\top\mathbf{K}^{-1}(\lambda_{B,i}\bar{\mathbf{x}}_{B,i} - \lambda_{B,j}\bar{\mathbf{x}}_{B,j}) = \mathbf{N}^\top(\mathbf{X}_{B,i} - \mathbf{X}_{B,j}).$$

Use the ground plane equation $\mathbf{N}^\top\mathbf{X}_{B,i} + \rho = 0$, and then the right-hand side of the equation above is 0. Also recall that we have defined $\mathbf{v} \triangleq \mathbf{KN}$ in the main paper, substitute $\mathbf{N} = \mathbf{K}^{-1}\mathbf{v}$ into the equation above, we have the following constraint

$$\mathbf{v}^\top\mathbf{K}^{-\top}\mathbf{K}^{-1}(\lambda_{B,i}\bar{\mathbf{x}}_{B,i} - \lambda_{B,j}\bar{\mathbf{x}}_{B,j}) = 0 \tag{2}$$

which is *scale-invariant* to $\lambda_{B,i}, \lambda_{B,j}$ and $\mathbf{v}$. As such, it is safe to substitute $\lambda_{B,i}, \lambda_{B,j}$ and $\mathbf{v}$ with their *scaled* version, *i.e.*, $\tilde{\lambda}_{B,i}, \tilde{\lambda}_{B,j}$ and $\tilde{\mathbf{v}}$ obtained in Section 3.2 of the main paper, leading to a *scale-invariant* linear constraint on $\mathbf{W} \triangleq \mathbf{K}^{-\top}\mathbf{K}^{-1} \in \mathbb{R}^{3\times3}$:

$$\tilde{\mathbf{v}}^\top\mathbf{W}(\tilde{\lambda}_{B,i}\bar{\mathbf{x}}_{B,i} - \tilde{\lambda}_{B,j}\bar{\mathbf{x}}_{B,j}) = 0. \tag{3}$$

A similar constraint can be derived for the shoulder center point of person i and j:

$$\tilde{\mathbf{v}}^\top\mathbf{W}(\tilde{\lambda}_{T,i}\bar{\mathbf{x}}_{T,i} - \tilde{\lambda}_{T,j}\bar{\mathbf{x}}_{T,j}) = 0. \tag{4}$$

which is essentially the same as Eq. (3) because

$$\begin{aligned}
&\tilde{\lambda}_{T,i}\bar{\mathbf{x}}_{T,i} - \tilde{\lambda}_{T,j}\bar{\mathbf{x}}_{T,j} \\
=&\mathbf{K}\mathbf{X}_{T,i} - \mathbf{K}\mathbf{X}_{T,j} \\
=&\mathbf{K}(\mathbf{X}_{B,i} + h\cdot\mathbf{N}) - \mathbf{K}(\mathbf{X}_{B,j} + h\cdot\mathbf{N}) \\
=&\mathbf{K}\mathbf{X}_{B,i} - \mathbf{K}\mathbf{X}_{B,j} \\
=&\tilde{\lambda}_{B,i}\bar{\mathbf{x}}_{B,i} - \tilde{\lambda}_{B,j}\bar{\mathbf{x}}_{B,j}.
\end{aligned}$$

Therefore, we construct the linear system $\mathbf{B}[1/f_x^2, 1/f_y^2]^\top = \mathbf{y}$ *only using the ankle* center points.

1.2 Linear System $\mathbf{B}[1/f_x^2, 1/f_y^2]^\top = \mathbf{y}$

The exact form of the second linear system $\mathbf{B}[1/f_x^2, 1/f_y^2] = \mathbf{y}$ is the following:

$$\mathbf{B} = \begin{bmatrix} \tilde{\mathbf{v}}^\top[1:2]\odot\Delta_{B,1,2}^\top[1:2] \\ \vdots \\ \tilde{\mathbf{v}}^\top[1:2]\odot\Delta_{B,i,j}^\top[1:2] \\ \vdots \\ \tilde{\mathbf{v}}^\top[1:2]\odot\Delta_{B,N-1,N}^\top[1:2] \end{bmatrix} \in \mathbb{R}^{\frac{N(N-1)}{2}\times2}, \qquad \mathbf{y} = \begin{bmatrix} \tilde{\mathbf{v}}[3]\Delta_{B,1,2}[3] \\ \vdots \\ \tilde{\mathbf{v}}[3]\Delta_{B,i,j}[3] \\ \vdots \\ \tilde{\mathbf{v}}[3]\Delta_{B,N-1,N}[3] \end{bmatrix} \in \mathbb{R}^{\frac{N(N-1)}{2}\times1} \tag{5}$$

where $\odot$ is component-wise product, we use $[a:b]$ to indicate the index range from a to b – both inclusive, and $[a]$ to indicate the a-th component of a vector. We also define $\Delta_{B,i,j} \triangleq \tilde{\lambda}_{B,i}\bar{\mathbf{x}}_{B,i} - \tilde{\lambda}_{B,j}\bar{\mathbf{x}}_{B,j}$.

We discuss the existence of solutions for both the *isotropic* and *anisotropic* focal length models. Note, in both cases, due to measurement noise, the solution of the least-square problem may not be positive – a requirement imposed by f_x^2 and f_y^2, and as such, the focal length may not be estimated given noisy measurements.

1.2.1 Anisotropic Focal Length $f_x \neq f_y$

By definition, at least one pair of people has to be observed to constrain $\mathbf{W}$, so $N >= 2$. When $N = 2$, $\mathbf{B}$ is of dimension 1×2 leading to an under-determined linear system which has infinitely many solutions. When $N >= 3$, $\mathbf{B}$ has more rows ($\frac{N(N-1)}{2} \geq 3$) than columns (2), and as such the linear system is over-determined – a solution exists in least-square sense. Therefore, we require *at least three people* visible to estimate f_x and f_y that are different.

1.2.2 Isotropic Focal Length $f_x = f_y$

We can simplify the intrinsics model even further by assuming isotropic focal length along x and y axes, *i.e.*, $f_x = f_y = f$, leading to the following least square estimate of f:

$$f^2 = -\frac{\sum_{i \neq j} \tilde{\mathbf{v}}[1:2]^\top \Delta_{B,i,j}[1:2]}{\sum_{i \neq j} \tilde{\mathbf{v}}[3] \Delta_{B,i,j}[3]} \tag{6}$$

As shown in the equation above, we need *at least one pair* of people ($i \neq j$)to solve f^2.

Results of noise-free simulation for both isotropic and anisotropic cases are reported in Table 1.

1.3 Reconstruction

In this Section, we derive the equations present in Sect. 3.3 *Reconstruction* of the main paper.

Ankle and shoulder center points in 3-D To reconstruct the ankle center points in 3-D, first recall the projection equation for the i-th ankle center point

$$\lambda_{B,i} \bar{\mathbf{x}}_{B,i} = \mathbf{K} \mathbf{X}_{B,i}.$$

It's trivial to compute $\mathbf{X}_{B,i}$ once the depth $\lambda_{B,i}$ and projection matrix $\mathbf{K}$ have been estimated – simply multiple both sides by $\mathbf{K}^{-1}$:

$$\mathbf{X}_{B,i} = \lambda_{B,i} \mathbf{K}^{-1} \bar{\mathbf{x}}_{B,i}.$$

The reconstruction equation for shoulder center points can be derived in the same way:

$$\mathbf{X}_{T,i} = \lambda_{T,i} \mathbf{K}^{-1} \bar{\mathbf{x}}_{T,i}.$$

Ground plane offset The ground plane equation for each ankle center point constrains the ground plane offset ρ:

$$\mathbf{N}^\top \mathbf{X}_{B,i} + \rho = 0.$$

Recall the ankle-shoulder relation $\mathbf{X}_{T,i} = \mathbf{X}_{B,i} + h \cdot \mathbf{N}$, and substitute $\mathbf{X}_{B,i}$ in the equation above with $\mathbf{X}_{T,i}$, we have another constraint for ρ:

$$\mathbf{N}^\top (\mathbf{X}_{T,i} - h \cdot \mathbf{N}) + \rho = 0.$$

As ρ is a scalar, the least-square estimate of ρ given all the constraints above is the average of the estimates from each individual constraint:

$$\begin{aligned} \rho &= \frac{1}{2N} \left(\sum_{i=1}^{N} -\mathbf{N}^\top \mathbf{X}_{B,i} + \sum_{i=1}^{N} h - \mathbf{N}^\top \mathbf{X}_{T,i} \right) \\ &= \frac{1}{2} h - \mathbf{N}^\top \frac{1}{2} (\bar{\mathbf{X}}_B + \bar{\mathbf{X}}_T) \end{aligned} \tag{7}$$

where $\bar{\mathbf{X}}_B = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_{B,i}$ and $\bar{\mathbf{X}}_T = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_{T,i}$.

Decoupled formulation Unlike the joint formulation where the depth $\lambda_{B,i}, \lambda_{T,i}$ and the calibration parameters $\mathbf{K}, \mathbf{N}, \rho$ are estimated simultaneously, in the decouple formulation, we reconstruct the 3-D points after estimating the calibration parameters. To reconstruct a 3-D point is to estimate its depth and back-project its 2-D observation to 3-D. To achieve this, recall the projection equation $\lambda_{B,i}\bar{\mathbf{x}}_{B,i} = \mathbf{K}\mathbf{X}_{B,i}$ and the ground plane equation $\mathbf{N}^\top\mathbf{X}_{B,i} + \rho = 0$, together we have a constraint on the unknown depth $\lambda_{B,i}$ and solve it as:

$$\lambda_{B,i} = -\frac{\rho}{\mathbf{N}^\top\mathbf{K}^{-1}\bar{\mathbf{x}}_{B,i}}. \tag{8}$$

We then back-project the ankle center point from 2-D to 3-D:

$$\begin{aligned}\mathbf{X}_{B,i} &= \lambda_{B,i}\mathbf{K}^{-1}\bar{\mathbf{x}}_{B,i} \\ &= \frac{-\rho}{\mathbf{N}^\top\mathbf{K}^{-1}\bar{\mathbf{x}}_{B,i}}\mathbf{K}^{-1}\bar{\mathbf{x}}_{B,i}.\end{aligned} \tag{9}$$

The corresponding shoulder center point can be computed as $\mathbf{X}_{T,i} = \mathbf{X}_{B,i} + h \cdot \mathbf{N}$.

1.4 Ground Plane Estimation Given Projection Matrix

In some use cases, the projection matrix $\mathbf{K}$ is given by a dedicated calibration procedure with a calibration pattern. In such cases, we only need to estimate the ground plane parameters $\mathbf{N}$ and ρ to reconstruct the 3-D points using the decoupled formulation. The physical distances are then measured within the 3-D reconstruction.

Let's start with the two projection equations for the i-th person:

$$\begin{aligned}\lambda_{T,i}\bar{\mathbf{x}}_{T,i} &= \mathbf{K}\mathbf{X}_{T,i} \\ \lambda_{B,i}\bar{\mathbf{x}}_{B,i} &= \mathbf{K}\mathbf{X}_{B,i}\end{aligned}$$

as the projection matrix $\mathbf{K}$ is known, we move it the left-hand side, and substitute $\mathbf{X}_{T,i} = \mathbf{X}_{B,i} + h \cdot \mathbf{N}$ leading to

$$\lambda_{T,i}\mathbf{K}^{-1}\bar{\mathbf{x}}_{T,i} - \lambda_{B,i}\mathbf{K}^{-1}\bar{\mathbf{x}}_{B,i} = h \cdot \mathbf{N} \tag{10}$$

which is linear in $\lambda_{T,i}, \lambda_{B,i}$ and $\mathbf{N}$. Collect and stack the constraints for all the people visible in the image, we have the following linear system to solve:

$$\begin{bmatrix} \mathbf{K}^{-1}\bar{\mathbf{x}}_{T,1} & -\mathbf{K}^{-1}\bar{\mathbf{x}}_{B,1} & \cdots & 0 & 0 & -h \cdot \mathbf{I}_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{K}^{-1}\bar{\mathbf{x}}_{T,N} & -\mathbf{K}^{-1}\bar{\mathbf{x}}_{B,N} & h \cdot \mathbf{I}_3 \end{bmatrix} \begin{bmatrix} \lambda_{H,1} \\ \lambda_{F,1} \\ \vdots \\ \lambda_{H,N} \\ \lambda_{F,N} \\ \mathbf{N} \end{bmatrix} = 0. \tag{11}$$

By solving this, we obtain $\tilde{\lambda}_H, \tilde{\lambda}_F$ and $\tilde{\mathbf{N}}$ up to scale. The proper scaling factor can be recovered by using the unitary constraint on the ground plane normal $\mathbf{N}$ as shown in the main paper. Estimation of the ground plane offset ρ as well as the 3-D points $\mathbf{X}$ is the same as discussed in Sect. 3.3 of the main paper.

1.5 Modeling Radial Distortion

The calibration algorithm described in the main paper assumes perfect perspective projection without considering lens distortion. Although in real-world experiments on MEVA, we have shown that our method is relatively robust to lens distortion, in this section, we extend our calibration algorithm by explicitly modeling lens distortion using the 1-parameter division model of Fitzgibbon [1].

Let the measured keypoints on the image be $\mathbf{x}' \in \Omega \subset \mathbb{R}^2$ which are distorted, and let their undistorted counterparts be $\mathbf{x} \in \Omega$, the 1-parameter division model relates the two via $\mathbf{x} = \mathbf{x}'/(1 + k \cdot r^2)$ where $r = \|\mathbf{x}'\| = \sqrt{x'^2 + y'^2}$, and $k \in \mathbb{R}$ is the unknown distortion parameter. Express the equality in homogeneous coordinates, we have

$$\bar{\mathbf{x}} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{1 + k \cdot r^2} \begin{bmatrix} x' \\ y' \\ 1 + kr^2 \end{bmatrix} = \frac{1}{1 + k \cdot r^2} \Big(\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} + k \begin{bmatrix} 0 \\ 0 \\ r^2 \end{bmatrix} \Big) = \frac{1}{1 + k \cdot r^2} (\bar{\mathbf{x}}' + k \cdot \mathbf{z}) \tag{12}$$

where we have defined $\mathbf{z} \triangleq [0, 0, r^2]^\top$.

Now if we substitute $\bar{\mathbf{x}}$ into the linear constraint Eq (1)

$$\lambda_{T,i} \bar{\mathbf{x}}_{T,i} - \lambda_{B,i} \bar{\mathbf{x}}_{B,i} = h \cdot \mathbf{v}$$

of the main paper, we have

$$\lambda'_{T,i}(\bar{\mathbf{x}}'_{T,i} + k\mathbf{z}_{T,i}) - \lambda'_{B,i}(\bar{\mathbf{x}}'_{B,i} + k\mathbf{z}_{B,i}) = h\mathbf{v} \tag{13}$$

where $\lambda' \triangleq \lambda/(1 + k \cdot r^2)$, and λ is the depth of the point. We construct $\bar{\mathbf{X}}'$ and $\mathbf{A}'$ by collecting and stacking all the constraints and unknowns. Along with an additional term $\mathbf{C} \in \mathbb{R}^{3N \times (2N+3)}$, we have a new system to solve

$$(\mathbf{A}' + k \cdot \mathbf{C})\bar{\mathbf{X}}' = \mathbf{0}. \tag{14}$$

Specifically, we have

$$\bar{\mathbf{X}}' = [\lambda'_{T,1}, \lambda'_{B,1} \cdots \lambda'_{T,N}, \lambda'_{B,N}, \mathbf{v}^\top]^\top \in \mathbb{R}^{2N+3}$$

$$\mathbf{A}' \triangleq \begin{bmatrix} \bar{\mathbf{x}}'_{T,1} & -\bar{\mathbf{x}}'_{B,1} & \cdots & 0 & 0 & -h \cdot \mathbf{I}_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \bar{\mathbf{x}}'_{T,N} & \bar{\mathbf{x}}'_{B,N} & -h \cdot \mathbf{I}_3 \end{bmatrix} \in \mathbb{R}^{3N \times (2N+3)}. \tag{15}$$

and

$$\mathbf{C} = \begin{bmatrix} \mathbf{z}_{T,1} & -\mathbf{z}_{B,1} & \cdots & 0_{3\times1} & 0_{3\times1} & 0_{3\times3} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0_{3\times1} & 0_{3\times1} & \cdots & \mathbf{z}_{T,N} & -\mathbf{z}_{B,N} & 0_{3\times3} \end{bmatrix} \in \mathbb{R}^{3N \times (2N+3)}. \tag{16}$$

Pre-multiply both sides of $(\mathbf{A}' + k \cdot \mathbf{C})\bar{\mathbf{X}}' = 0$ by $\mathbf{A}'^\top$, we obtain a generalized eigenvalue problem $(\mathbf{A}'^\top \mathbf{A}')\bar{\mathbf{X}}' = k \cdot (-\mathbf{A}'^\top \mathbf{C})\bar{\mathbf{X}}'$, which can be solved by QZ decomposition [4].

The standard form of a generalized eigenvalue problem is $\mathbf{Au} = \lambda\mathbf{Bu}$, where both $\mathbf{A}$ and $\mathbf{B}$ are square matrices. In our case, $\mathbf{A} \coloneqq \mathbf{A}'^\top \mathbf{A}'$, $\mathbf{B} \coloneqq -\mathbf{A}'^\top \mathbf{C}$, $\mathbf{u} \coloneqq \bar{\mathbf{X}}'$, and $k \coloneqq \lambda$.

We show simulation results of distortion modeling in Sect. 2.2 and compare it against the vanilla version of our estimator that *does not* model lens distortion.

2 More Simulation Results

2.1 Noise-Free

We report detailed experiment results of the noise-free simulation for both the isotropic ($f_x = f_y$) and anisotropic ($f_x \neq f_y$) focal length model in Table 1. The experiment setup and the evaluation metrics are described in the main paper. We found that all the error terms are close to 0 (to machine precision) confirming the validity of the proposed algorithm.

Error (unit)	FOV \res.	640 × 480	1280 × 720	1920 × 1080
isotropic focal length $f_x = f_y$				
f(%)	45°	1.89e-11	1.43e-11	8.02e-11
	60°	1.38e-11	5.66e-11	2.87e-11
	90°	4.19e-11	1.18e-10	7.33e-11
	120°	2.58e-11	7.36e-11	4.15e-10
$\mathbf{N}$(°)	45°	9.04e-8	9.41e-8	1.03e-7
	60°	9.45e-8	9.48e-8	9.99e-8
	90°	1.01e-7	8.90e-8	9.73e-8
	120°	8.94e-8	9.91e-8	9.38e-8
ρ(%)	45°	7.96e-12	6.85e-12	2.18e-11
	60°	7.82e-12	2.60e-11	1.27e-11
	90°	1.80e-11	1.62e-11	2.56e-11
	120°	9.33e-12	2.03e-11	2.98e-11
$\mathbf{X}$(%)	45°	1.11e-11	9.65e-12	4.41e-11
	60°	9.42e-12	3.47e-11	1.70e-11
	90°	2.94e-11	6.05e-11	3.92e-11
	120°	1.46e-11	3.75e-11	2.14e-10
anisotropic focal length $f_x \neq f_y$				
f_x(%)	45°	2.34e-11	5.85e-11	4.68e-11
	60°	1.31e-11	4.23e-11	1.76e-11
	90°	1.69e-11	1.76e-11	4.27e-11
	120°	6.59e-11	4.51e-11	4.11e-11
f_y(%)	45°	1.36e-11	3.98e-11	2.68e-11
	60°	1.19e-11	1.58e-11	2.63e-11
	90°	1.82e-11	2.31e-11	1.91e-11
	120°	4.45e-11	4.01e-11	2.97e-11
$\mathbf{N}$(°)	45°	9.68e-8	9.31e-8	1.02e-7
	60°	8.99e-8	8.75e-8	8.97e-8
	90°	8.99e-8	9.95e-8	9.97e-8
	120°	1.01e-7	9.19e-8	9.06e-8
ρ(%)	45°	2.97e-12	2.76e-12	4.08e-12
	60°	2.69e-12	4.36e-12	5.09e-12
	90°	3.38e-12	4.27e-12	3.90e-12
	120°	3.55e-12	8.34e-12	4.88e-12
$\mathbf{X}$(%)	45°	7.47e-12	2.33e-11	2.22e-11
	60°	6.24e-12	1.78e-11	1.55e-11
	90°	1.08e-11	1.36e-11	2.69e-11
	120°	3.90e-11	2.79e-11	3.20e-11

Table 1: *Estimation error in noise-free simulation.* We conduct Monte Carlo experiments of 5,000 trials for each resolution-FOV pair, and show the average of focal length estimation error (f in isotropic case and f_x, f_y in anisotropic case), ground plane estimation error ($\mathbf{N}, \rho$), and reconstruction error ($\mathbf{X}$) as described in Sect. 4 *Simulation* of the main paper.

2.2 Lens Distortion

In this section, we demonstrate the proposed calibration method with distortion modeling in simulation.

The simulation has a similar setup as in the main paper where we first randomly generate ankle

and shoulder center points $\mathbf{X}_{B,i}, \mathbf{X}_{T,i} \in \mathbb{R}^3$ in 3-D satisfying the three model assumptions, and then project the 3-D points to the image plane to produce 2-D measurements $\mathbf{x}_{B,i}, \mathbf{x}_{T,i} \in \Omega$.

Unlike the perfect perspective projection model in the main paper where no lens distortion is applied, in this experiment, we adopt a polynomial model [1] to synthesize the distorted measurements: $\mathbf{x}_d = \mathbf{c} + (1 + k_1 \cdot r^2 + k_2 \cdot r^4) \cdot (\mathbf{x} - \mathbf{c}) \in \mathbb{R}^2$, where $\mathbf{x}$ is the undistorted measurement, $\mathbf{c}$ is the principal point, $r = \|\mathbf{x} - \mathbf{c}\|$ is the distance of the undistorted projection to the principal point, and k_1, k_2 are the distortion parameters. Note, the polynomial model used in synthesizing the simulation data is different from the 1-parameter division model used in the solver. Though the polynomial model can also be used in modeling, we found the 1-parameter division model results in a simpler implementation where no iterative optimization is needed.

We fix the resolution to 1920×1080, FOV to $90°$, number of people visible in the image to 20, and test different distortion configurations. For each configuration, we conduct Monte Carlo experiments of 5,000 trials. Table 2 shows a comparison of our estimator with and without modeling lens distortion on the *noise-free but distorted measurements* under different distortion configurations (k_1, k_2). It's not hard to see that in all the test cases, the estimator that models lens distortion has less estimation error compared to the one that does not model lens distortion. However, we also found that the former is more sensitive to measurement noise than the latter, and fails more often in noisy settings. We leave the seek of different distortion models and more stable numeric schemes as future work.

[1] https://docs.opencv.org/3.4/d4/d94/tutorial_camera_calibration.html

Error (unit)	with distortion modeling	without distortion modeling
$k_1 = 10^{-3}, k_2 = 0$		
$f_x(\%)$	$\mathbf{1.077e-5 \pm 7.81e-6}$	0.16 ± 0.32
$f_y(\%)$	$\mathbf{1.077e-5 \pm 7.81e-6}$	0.16 ± 0.32
$\mathbf{N}(^\circ)$	$\mathbf{8.99e-7 \pm 7.47e-7}$	0.053 ± 0.045
$\rho(\%)$	$\mathbf{3.55e-6 \pm 5.29e-6}$	0.12 ± 0.12
$\mathbf{X}(\%)$	$\mathbf{6.06e-6 \pm 3.46e-6}$	0.23 ± 0.23
$k_1 = -10^{-3}, k_2 = 0$		
$f_x(\%)$	$\mathbf{1.076e-5 \pm 7.78e-6}$	0.17 ± 0.34
$f_y(\%)$	$\mathbf{1.076e-5 \pm 7.78e-6}$	0.17 ± 0.34
$\mathbf{N}(^\circ)$	$\mathbf{1.049e-6 \pm 2.90e-7}$	0.054 ± 0.046
$\rho(\%)$	$\mathbf{3.68e-6 \pm 5.25e-6}$	0.13 ± 0.21
$\mathbf{X}(\%)$	$\mathbf{6.09e-6 \pm 3.39e-6}$	0.24 ± 0.33
$k_1 = 10^{-4}, k_2 = 0$		
$f_x(\%)$	$\mathbf{3.82e-6 \pm 8.64e-6}$	0.017 ± 0.055
$f_y(\%)$	$\mathbf{3.82e-6 \pm 8.64e-6}$	0.017 ± 0.055
$\mathbf{N}(^\circ)$	$\mathbf{1.44e-6 \pm 3.80e-6}$	0.0055 ± 0.013
$\rho(\%)$	$\mathbf{2.71e-6 \pm 8.52e-6}$	0.013 ± 0.043
$\mathbf{X}(\%)$	$\mathbf{3.86e-6 \pm 6.82e-6}$	0.023 ± 0.045
$k_1 = -10^{-4}, k_2 = 0$		
$f_x(\%)$	$\mathbf{3.72e-6 \pm 8.24e-6}$	0.017 ± 0.035
$f_y(\%)$	$\mathbf{3.72e-6 \pm 8.24e-6}$	0.017 ± 0.035
$\mathbf{N}(^\circ)$	$\mathbf{1.41e-6 \pm 3.56e-6}$	0.0053 ± 0.0059
$\rho(\%)$	$\mathbf{2.70e-6 \pm 8.54e-6}$	0.013 ± 0.026
$\mathbf{X}(\%)$	$\mathbf{3.82e-6 \pm 6.54e-6}$	0.023 ± 0.034
$k_1 = 10^{-4}, k_2 = 10^{-5}$		
$f_x(\%)$	$\mathbf{0.00079 \pm 0.0013}$	0.047 ± 0.27
$f_y(\%)$	$\mathbf{0.00079 \pm 0.0013}$	0.047 ± 0.27
$\mathbf{N}(^\circ)$	$\mathbf{0.00035 \pm 0.00057}$	0.012 ± 0.068
$\rho(\%)$	$\mathbf{0.00054 \pm 0.0013}$	0.029 ± 0.19
$\mathbf{X}(\%)$	$\mathbf{0.00089 \pm 0.0011}$	0.053 ± 0.18
$k_1 = -10^{-4}, k_2 = 10^{-5}$		
$f_x(\%)$	$\mathbf{0.0019 \pm 0.0036}$	0.020 ± 0.051
$f_y(\%)$	$\mathbf{0.0019 \pm 0.0036}$	0.020 ± 0.051
$\mathbf{N}(^\circ)$	$\mathbf{0.00058 \pm 0.0018}$	0.0039 ± 0.013
$\rho(\%)$	$\mathbf{0.00097 \pm 0.0037}$	0.0067 ± 0.044
$\mathbf{X}(\%)$	$\mathbf{0.0016 \pm 0.0031}$	0.015 ± 0.042

Table 2: *Estimation error with and without modeling lens distortion* under different distortion configurations (k_1, k_2). We show $mean \pm std$ of the various estimation errors of the two estimators, and highlight the best in **bold**.

3 Sitting People Classifier

In the conclusion of the main paper, we mentioned that the estimator can be further improved by filtering out people that are not upright such as sitting people. While this can be done by training a neural network, we show here that a simple sitting people classifier can be built using keypoint information only.

The HRNet [8] pose detector we adopt in the main paper predicts 17 2-D human body keypoints for each person. Let $\mathbf{x}_{j,i} \in \mathbb{R}^2, j \in \mathcal{J} \triangleq \{1 \cdots 17\}, i = 1 \cdots N$, where $\mathcal{J}$ is the index set of the keypoints, and N is the number of people in the image. We base our classifier on the following heuristic: When we use a minimal enclosing ellipse to cover the keypoints detected on a person, the shape of the ellipse is elongated when the person is standing, whereas for a sitting person, the shape of the ellipse approaches a circle. Mathematically, we compute the ratio of the two eigenvalues of the covariance matrix of the 17 keypoints as an indicator of *how elongated the ellipse is*, *i.e.*, larger ratio means more likely that the person is standing. The covariance matrix of the i-th person's

keypoints reads

$$\mathbf{cov}_i = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} (\mathbf{x}_{j,i} - \bar{\mathbf{x}}_i)(\mathbf{x}_{j,i} - \bar{\mathbf{x}}_i)^\top \tag{17}$$

where $\bar{\mathbf{x}}_i = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \mathbf{x}_{j,i}$ is the centroid of the 2-D keypoints on the i-th person.

For an overhead camera from which the keypoints detected on a person concentrate into a small cluster – more specifically, the ankle and shoulder center point of the person approximately coincide, the distribution of keypoints is *not* a good indicator of sitting/standing people. Fortunately, with wide Field-of-View (FOV) cameras, we expect that even for overhead configuration, the ankle and shoulder center point pairs appearing on the *peripheral* of the image do not degrade into *a single dot* and thus inform the scale of the scene and the focal length of the camera.

4 More Results on MEVA

4.1 More Visualization on MEVADA

Fig. 1 shows more qualitative results on the MEVADA test set, where in each plot, the estimated ground plane is superimposed on the RGB image as cyan regular grids (each grid is $2m \times 2m$). Bounding boxes and keypoints are shown for the pair of selected people of which the distance is to be estimated. We show a link connecting the ankle center point of the pair of people with the estimated distance (in meters) marked on the link. For the bounding boxes, keypoints, and links, red indicates the two people are less than 2 meters apart and green otherwise. Estimated distance, as well as the ground truth label collected from human reviewers, are displayed at the top-left corner of each image.

We found that our system is able to estimate the ground plane and distances reasonably well for both indoor and outdoor scenarios. Estimated distances are consistent with human labels for each image and are reasonable with closer examination. Fig 1a, 1b, 1g are examples where people are fairly close. The model can pick the information up on both ground planes or area with small elevation such as staircases. In Fig 1c, the distance between the selected people is further than a sedan, which is typically about 5 meters long. In Fig 1e, the distance from the sideline to the middle of the court is 7.6 meters that is half of the width of a basketball court. That justifies the 11.9 meter distance estimate which is about 3/4 of the court width.

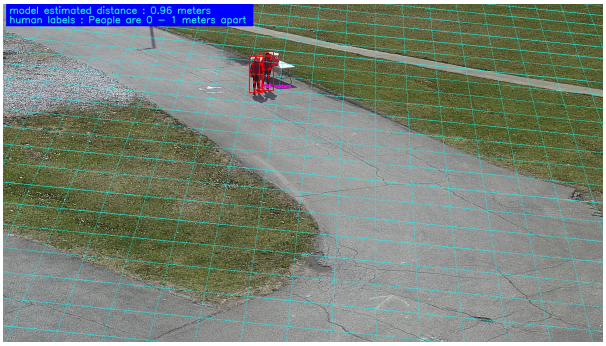
4.2 Detailed Calibration Results on MEVA

Detailed focal length estimation error on MEVA dataset can be found below:

Video	fx error (%)	fy error (%)
2018-03-09.09-06-41.09-10-01.hospital.G301.krtd	4.02	2.92
2018-03-05.13-30-00.13-35-00.bus.G340.krtd	3.57	5.14
2018-03-05.13-25-01.13-30-01.school.G328.krtd	6.67	1.24
2018-03-05.15-55-00.16-00-00.hospital.G341.krtd	6.77	4.56
2018-03-15.08-34-31.08-35-00.school.G299.krtd	10.01	9.91
2018-03-05.13-20-00.13-25-00.hospital.G436.krtd	11.23	10.17
2018-03-07.11-45-09.11-50-09.bus.G340.krtd	12.36	11.55
2018-03-16.08-31-36.08-35-00.school.G330.krtd	13.51	13.59
2018-03-06.15-05-02.15-10-02.school.G336.krtd	16.81	14.00
2018-03-05.10-50-00.10-55-00.hospital.G301.krtd	20.42	12.27
2018-03-05.11-20-00.11-25-00.bus.G340.krtd	22.84	21.71
2018-03-05.14-05-00.14-10-00.hospital.G341.krtd	27.39	25.53
2018-03-07.13-15-01.13-20-01.school.G638.krtd	28.81	29.28
2018-03-05.13-20-01.13-25-00.bus.G505.krtd	20.68	31.19
2018-03-05.14-15-00.14-20-00.hospital.G301.krtd	24.30	31.76
2018-03-05.18-10-00.18-15-00.hospital.G436.krtd	20.43	34.56
2018-03-05.12-25-00.12-30-00.school.G336.krtd	31.78	36.33
2018-03-05.11-15-00.11-20-00.school.G339.krtd	2.57	39.73
2018-03-07.11-35-09.11-40-09.bus.G340.krtd	43.32	43.28
2018-03-05.14-20-00.14-25-00.school.G339.krtd	43.85	37.12
2018-03-07.17-25-03.17-30-03.school.G300.krtd	25.16	46.53
2018-03-05.16-50-00.16-55-00.bus.G509.krtd	46.88	46.88
2018-03-05.13-15-00.13-20-00.bus.G506.krtd	6.45	49.52
2018-03-05.18-25-00.18-29-31.school.G424.krtd	53.60	37.45
2018-03-05.09-50-07.09-55-00.school.G300.krtd	29.69	66.12
2018-03-14.16-20-01.16-25-01.school.G639.krtd	70.27	0.44

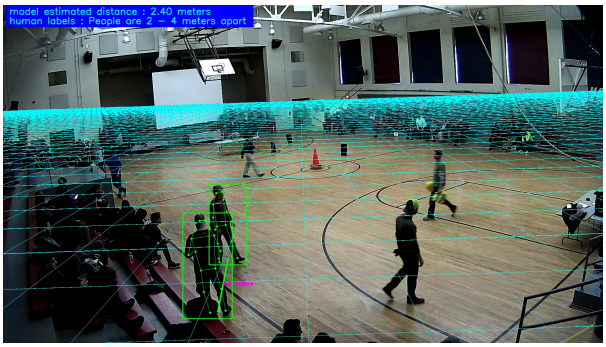
Table 3: Focal length estimation error.

(a) *outdoor - ground plane*



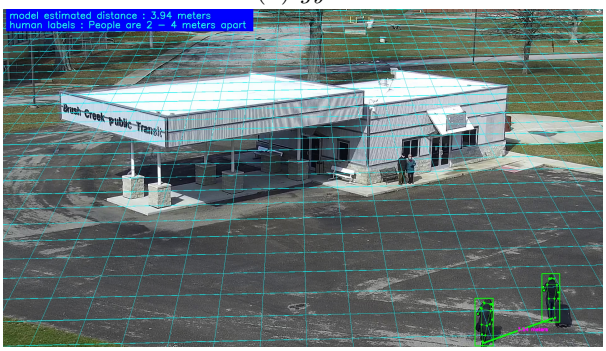
(b) *outdoor - ground plane*

(c) *outdoor - ground plane*

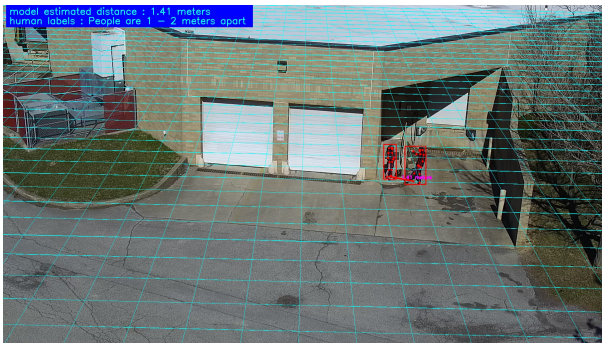


(d) *gym*

(e) *gym*



(f) *outdoor - ground plane*



(g) *outdoor - small elevation*

Figure 1: More qualitative results on MEVADA.

5 URL of the Public Datasets Used in Paper

- Oxford Town Center [5] https://megapixels.cc/oxford_town_centre/
- MEVA [6]: https://mevadata.org/#getting-data
- vPTZ [7]: https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/vptz/
- POM [2]: https://www.epfl.ch/labs/cvlab/data/data-pom-index-php/

6 Demo Video

The attached video `OxfordTownCenterDemo.mov` shows a video demo of our end-to-end system running on the publicly available Oxford Town Center dataset [5]. Fig. 1 of the main paper is taken from the demo video.

In the **left** panel, we show the keypoints of each detected person in two colors where red means the person is within 6 feet from others – potentially *unsafe* according to social distancing guidelines [3], and green means the person is at least 6 feet from others and *safe*. We estimate the projection matrix $\mathbf{K}$ and ground plane parameters $\mathbf{N}, \rho$ using all the keypoints detected in the video in batch mode and visualize the ground plane as the regular grids in cyan overlaid on the video frames. Each grid cell is $6ft. \times 6ft.$ The calibration parameters are then used in the *decoupled formulation* to reconstruct the 3-D coordinates of the ankle center points from which the distances are measured. The estimated metric distances (in feet) are superimposed (in pink) on the nearest neighbor (shown as the green/red link) of each person. Green links mean safe and red unsafe.

In the **top right** panel, we show a top-down view of the scene, where each dot represents a person, and red means unsafe and green safe.

In the **bottom right** panel, we show a heat map of individuals considered unsafe (within 6 feet from others) aggregated over time, brighter means higher density, and darker lower density. The heat map can be used to guide potential safety measures to be taken, for instance, workplace re-arrangement.

References

[1] Andrew W Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 5

[2] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):267–282, 2007. 12

[3] Centers for Disease Control and Prevention. Social distancing – keep a safe distance to slow the spread. "https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html". 12

[4] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012. 5

[5] Jules. Harvey, Adam. LaPlace. Megapixels: Origins and endpoints of datasets created "in the wild", 2019-2020. 12

[6] Kitware. Meva: Multiview extended video with activities. "https://mevadata.org/". 12

[7] Horst Possegger, Matthias Rüther, Sabine Sternig, Thomas Mauthner, Manfred Klopschitz, Peter M. Roth, and Horst Bischof. Unsupervised Calibration of Camera Networks and Virtual PTZ Cameras. In *Proc. Computer Vision Winter Workshop (CVWW)*, 2012. 12

[8] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 8