Supplementary

Stacked Homography Transformations for Multi-View Pedestrian Detection

Liangchen Song[1], Jialian Wu[1], Ming Yang[2], Qian Zhang[2], Yuan Li[3], and Junsong Yuan[1]
[1]University at Buffalo [2]Horizon Robotics, Inc. [3]Google, Inc.
{lsong8,jialianw,jsyuan}@buffalo.edu

1. Full proof of Proposition 1

An intuitive understanding about the proposition is that with two notations we are able to determine the vertical vanishing point of the camera.

Proposition 1. *If* $\mathbf{K}, \mathbf{K}_g$ *and* $\mathbf{H}_0$ *are known, we can construct a stack of transformations* $\{(\mathbf{H}_0, \cdots, \mathbf{H}_D)\}$ *with two extra annotations of pedestrians, where the two pedestrians should have the same height. Two extra annotations means two sets of points correspondences in the camera image and BEV image,* i.e.*,* $(\boldsymbol{f}_1, \boldsymbol{h}_1, \boldsymbol{o}_1)$ *and* $(\boldsymbol{f}_2, \boldsymbol{h}_2, \boldsymbol{o}_2)$.

Proof. First recall that we denote the extrinsic matrix as $\mathbf{E}^i = [\mathbf{R}^i|\mathbf{t}^i] = (\mathbf{e}_1^i, \mathbf{e}_2^i, \mathbf{e}_3^i, \mathbf{e}_4^i)$, where each $\mathbf{e}$ is a column vector, the matrix for the projection from $Z = 0$ plane to the screen plane is $\mathbf{K}^i\mathbf{E}_0^i$ where $\mathbf{E}_0^i = (\mathbf{e}_1^i, \mathbf{e}_2^i, \mathbf{e}_4^i)$, as the world coordinates will be $\mathbf{e}_1^i X + \mathbf{e}_2^i Y + \mathbf{e}_4^i$. Therefore, we can recover part of extrinsic parameters with $\mathbf{E}_0 = \mathbf{K}^{-1}\mathbf{H}_0^{-1}\mathbf{K}_g$, then if we define $\mathbf{E}_D = \mathbf{K}^{-1}\mathbf{H}_D^{-1}\mathbf{K}_g$, we have $\mathbf{E}_D = \mathbf{E}_0 + \Delta\mathbf{T}$ where $\Delta\mathbf{T}$ is

$$\begin{pmatrix} 0 & 0 & \Delta t_1 \\ 0 & 0 & \Delta t_2 \\ 0 & 0 & \Delta t_3 \end{pmatrix}. \tag{1}$$

To construct a set of transformations, we only need to know $\Delta t_1, \Delta t_2$ and Δt_3. Next, from one annotation $(\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{o})$ we have the equations

$$\begin{cases} \mathbf{E}_0\mathbf{K}_g^{-1}\boldsymbol{o} \sim \mathbf{K}^{-1}\boldsymbol{f}, \\ \mathbf{E}_D\mathbf{K}_g^{-1}\boldsymbol{o} \sim \mathbf{K}^{-1}\boldsymbol{h}. \end{cases} \tag{2}$$

Next, we denote that h_u, h_v, f_u, f_v and o_x, o_y are the first and second elements from $\mathbf{K}^{-1}\boldsymbol{h}$, $\mathbf{K}^{-1}\boldsymbol{f}$ and $\mathbf{K}_g^{-1}\boldsymbol{o}$, respectively. e_{31}, e_{32} and e_{33} are elements of the third row of $\mathbf{E}_0$. Equation (2) can be rewritten as

$$\begin{cases} \dfrac{e_{11}o_x + e_{12}o_y + e_{13}}{e_{31}o_x + e_{32}o_y + e_{33}} = f_u \\ \dfrac{e_{21}o_x + e_{22}o_y + e_{23}}{e_{31}o_x + e_{32}o_y + e_{33}} = f_v \\ \dfrac{e_{11}o_x + e_{12}o_y + e_{13} + \Delta t_1}{e_{31}o_x + e_{32}o_y + e_{33} + \Delta t_3} = h_u \\ \dfrac{e_{21}o_x + e_{22}o_y + e_{23} + \Delta t_2}{e_{31}o_x + e_{32}o_y + e_{33} + \Delta t_3} = h_v \end{cases} \tag{3}$$

These equations can be converted to

$$\begin{cases} f_u + \dfrac{\Delta t_1}{e_{31}o_x + e_{32}o_y + e_{33}} = h_u(1 + \dfrac{\Delta t_3}{e_{31}o_x + e_{32}o_y + e_{33}}) \\ f_v + \dfrac{\Delta t_2}{e_{31}o_x + e_{32}o_y + e_{33}} = h_v(1 + \dfrac{\Delta t_3}{e_{31}o_x + e_{32}o_y + e_{33}}) \end{cases} \tag{4}$$

So from Equation (3), we have two equations in total for solving $\Delta\mathbf{T}$,

$$\begin{cases} \Delta t_1 - h_u \Delta t_3 = (h_u - f_u)(e_{31}o_x + e_{32}o_y + e_{33}), \\ \Delta t_2 - h_v \Delta t_3 = (h_v - f_v)(e_{31}o_x + e_{32}o_y + e_{33}), \end{cases} \tag{5}$$

As each point provides two equations and there are three variables, we need two extra annotations for constructing the transformations. □

2. Additional results

2.1. Training with the estimated homographies

We mathematically analyze the correctness of the propositions based on the assumptions and then Fig.9 validates the assumptions with real data. We conduct detection experiments as suggested: MODA results on MultiviewX (repeated 10 times) are listed in the table below. The difference between the two settings is statistically negligible.

	mean	std	min	max	median
Given extrinsics	88.27	0.10	87.98	88.42	88.30
Estimated extrinsics via propositions	88.25	0.11	88.01	88.43	88.27