Automatic Calibration 'In the Wild' of Monocular Surveillance Camera Solely from Pedestrian Height

Martín E. Oviedo and Sebastián I. Arroyo

August 2, 2020

Automatic Calibration ‘in the wild’ of Monocular Surveillance Camera solely from Pedestrian Height

1st Martín E. Oviedo
Carrera de Ingenieria en Automatización Y Control Industrial
Universidad Nacional de Quilmes
Bernal, Argentina
eliasoviedo1718@gmail.com

2nd Sebastián I. Arroyo
Departamento de Ciencia y Tecnología
Universidad Nacional de Quilmes
Bernal, Argentina
sebastian.arroyo@unq.edu.ar

***Abstract*—We propose a method that relies solely on pedestrian detections in the wild (no known fiducial points) and uses their average height as measuring stick to calibrate monocular surveillance systems. The core idea is that pedestrians feet and head are aligned vertically and the calibration parameters must therefore predict back-projected positions of the feet and head that have the same horizontal coordinates. The advantages are that it does not require any fiducial points; nor to assume that the motion of pedestrians is aligned with vanishing points in the scene; nor relies on a stable framerate as it works on static images. The results are promising: the predicted world positions of pedestrians minimise the vertical misalignment of head and feet and they with the truth positions as well as expected. The approach is very simple and the results encourage us to develop it further with a bayesian approach.**

***Index Terms*—imaging, calibration, automatic, pedestrian**

I. Introduction

The ubiquitous presence of surveillance cameras and the proliferation of pedestrian detection algorithms [1] leads to the obvious question of how to geolocate the pedestrians detected in the video. The method of choice is a perspective projection [2] where the 2D image detections are transformed via a homography to the 2D ground plane. The question now becomes how to estimate the homography matrix needed. The difficulty resides in doing so for surveillance cameras ‘in the wild’, i.e. without additional information other than the video feed (camera make and model, optical distortion parameters, installation position, objects of known world position, etc). In this work we present a very simple framework that uses solely pedestrian detections to estimate the projection matrix.

Reference [3] tackles the problem by using multiple cameras, they state one of the core ideas in automatic calibration from pedestrian detections: “The desired information from the pedestrian is the position in each image of the top of their head and the point on the ground directly beneath. When a person is standing, the point on the ground directly beneath their head will be where their feet touch the ground”. Facing the single view problem, [4] combines a pedestrian detection algorithm together with the homography estimation. They demonstrate that human body can be used as a calibration target by observing a person standing at various locations. In [5] the motion of pedestrians and vehicles is used to estimate the vanishing points. This information is combined with the height of the camera above the ground (which is assumed known) to estimate the intrinsic and extrinsic parameters of the camera. An interesting approach is taken in [6] where the measuring stick is the velocity vector of pedestrians in motion, the velocity is assumed constant thruout the motion. It is aimed at crowded scenes where the view of pedestrians by the camera is guaranteed to be partially occluded.

In this work we exploit, firstly, the property already stated by [3] that pedestrians heads and feet have the same x, y in world coordinates. And secondly, that a minimal parametrisation of the pin hole model yields two homography matrices: one that back-projects the feet of pedestrians from the image to the ground plane, and a second one that back-projects the head of pedestrians to the horizontal plane at $z = h$, the ‘heads’ plane. The only required input data is several detections of the heads and feet of pedestrians and to provide the value of h, the average physical height of pedestrians. There is no need for multiple cameras as in [3], or the reliance on a particular pedestrian detection algorithm as proposed in [4] (arguably one might want to switch to other detection methods or to use a different object other than pedestrians). Surveillance cameras could be installed to observe any kind of scene, so it is not guaranteed that the motion of pedestrians would be of use to determine vanishing points as in [5]. Finally, the calculation of velocities in [6] requires some degree of reliability of the timestamps of the frames, or stability of the communication network that streams the video, we aim to solve the problem from static captures of the scene.

This work is very similar to [7]. They only require the top and bottom detections of a target object standing at various locations in the scene. The difference is only in the specific mathematical solution proposed, here we aim to simplify the implementation by relying in optimization routines.

A. The gist: back-projection of the head must be above the back-projection of the feet

We minimalistically formalise the pin hole image formation model as parameterised by only four magnitudes: focal distance, two orientation angles and height of the camera above

Universidad Nacional de Quilmes, project *Sistemas Autónomos Basados en Inteligenica Artificial*, under *Programa I+D UNQ EX- PTE 1303/19*. Science and Technology Department, Industrial Automation and Control Engineering.

the ground. This parameters determine a pin hole projection matrix. Which in turn is used to construct the homography projection matrices that map image coordinates in pixels to world points on the ground or on the $z = h$ plane. The x, y coordinates of both mappings must be equal because a pedestrians head is directly above the feet. We propose that the error function to be optimised is the quadratic error between them.

II. Model Definition with Reduced Degrees of Freedom

The pinhole camera model [8]–[10] maps a 3D coordinate (x, y, z) in the world frame of reference to a 2D image coordinate (u, v) like so:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} [R \mid T] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{1}$$

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K [R \mid T] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = M \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{2}$$

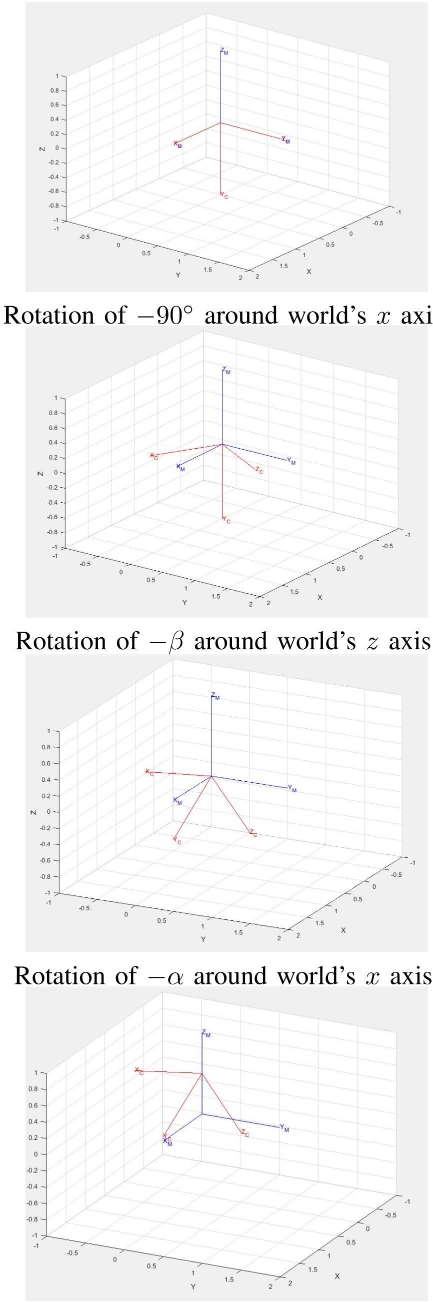
where the rotation matrix $R^{3\times3}$ has three degrees of freedom (interpreted as angles as in the Euler convention or components of a Rodrigues vector) and the 3D translation vector $T^{3\times1}$ has another three degrees of freedom. The intrinsic projection matrix K is defined with three parameters. In total we would need to estimate nine parameters to define the projection. We argue that we can simplify the problem to much fewer parameters, in this work we will only need to estimate four parameters. To begin, the intrinsic parameters (c_x, c_y) can be assumed to be at the exact center of the image [3], leaving only f to be estimated.

Arguably, out of the six roto translation parameters only three can be estimated from the detections of pedestrians in the image: the orientation of the ground plane with respect to the camera and the distance between them. The data that is available is the pixel coordinates of the heads and feet of pedestrians, their corresponding world positions are unknown. Without loss of generality, the origin of the ground plane is arbitrarily defined to be below the camera, and the y versors of the world and camera frame of reference are contained in the same vertical plane. We chose to parameterize the roto-translation in terms of fixed axis x-z-x rotations as

$$[R \mid T] = \left[\begin{array}{c|c} & 0 \\ R_x(-\alpha) R_z(-\beta) R_x(-90) & 0 \\ & t \end{array}\right]^{-1} \tag{3}$$

that are easy to interpret as shown in Fig. 1, assuming the camera to be directly above the origin of the world t is the height of the camera above the ground, and α, β are angles of rotation.

The selected minimalistic parameters f, α, β, t univocally define M that projects any 3D world coordinates into the image by the pin hole model in (2). In what follows we will exploit two particular cases: world coordinates restricted to the ground plane and restricted to the horyzontal $z = h$ plane.



Rotation of $-90°$ around world's x axis.

Rotation of $-\beta$ around world's z axis.

Rotation of $-\alpha$ around world's x axis.

Translation of t along the world's z axis.

Fig. 1. The transformations that define $[R \mid T]^{-1}$

A. Back projecting feet and head to calculate error function

Pedestrians feet are in the ground plane and pedestrian heads are in a horizontal plane at $z = h$ above the ground. The perspective transform that projects those world horizontal planes unto the image are defined by homography matrices H_f and H_h, we will calculate them from h and the four columns of M denoted as M_i. Starting with the feet, their world coordinate has $z = 0$ so it is evident that

$$H_f = [M_1 \mid M_2 \mid M_4]. \tag{4}$$

The projection for the heads can be rearranged as

$$s_h \begin{bmatrix} u_h \\ v_h \\ 1 \end{bmatrix} = [M_1 \mid M_2 \mid hM_3 + M_4] \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{5}$$

$$\text{so evidently } H_h = [M_1 \mid M_2 \mid hM_3 + M_4]. \tag{6}$$

A pedestrian's feet detected in the image at (u_f, v_f) must yield the same (x, y) when back projected to the ground plane than the head detected at (u_h, v_h) when back projected to the $z = h$ plane. For some set of values (f, α, β, t) the back projection error is calculated as follows. Fist, calculate M as defined in (2) and (3). Combine its columns and the constant h to calculate the homography matrices H_F and H_h as in (4) and (6).

Back project the detections of feet and head to their corresponding horizontal world planes by doing

$$\begin{bmatrix} x_f \\ y_f \\ 1 \end{bmatrix} = \frac{1}{s_f} H_f^{-1} \begin{bmatrix} u_f \\ v_f \\ 1 \end{bmatrix} \tag{7}$$

$$\text{and} \begin{bmatrix} x_h \\ y_h \\ 1 \end{bmatrix} = \frac{1}{s_h} H_h^{-1} \begin{bmatrix} u_h \\ v_h \\ 1 \end{bmatrix}. \tag{8}$$

And finally calculate the squared error as

$$E = (x_f - x_h)^2 + (y_h - y_f)^2. \tag{9}$$

The sum over all the data has been omitted for simplicity.

Starting from an ad-hoc defined seed value for f, α, β, t we use a non linear minimization routine in scipy [11] to find the point estimates of the most likely values of the parameters.

III. Results

To test the method we took a picture with 16 objects of equal height ($h = 12$cm) to simulate pedestrians. To validate our method we generated ground truth world positions of the objects by marking four 'tag' points whose world position were measured by hand. This tag points were used to estimate a homography matrix that maps the feet of the detected pedestrians to their true world position. Fig. 2 shows the detections of simulated pedestrians (both head and feet) along with the four tag points. The true world positions of the pedestrians is shown in Fig. 3 as blue dots, the red dots are the tag points.

The seed values were

$$(f, \alpha, \beta, t) = (c_x, 36^\circ, 36^\circ, 7h).$$

that correspond to a quadratic error as defined in (9) of $E_{\text{seed}} = 1374679.829\text{cm}^2$. The optimization routine returned the optimal values of

$$(f, \alpha, \beta, t) = (2854.24, 2.54^\circ, 31.81^\circ, 64.9cm)$$

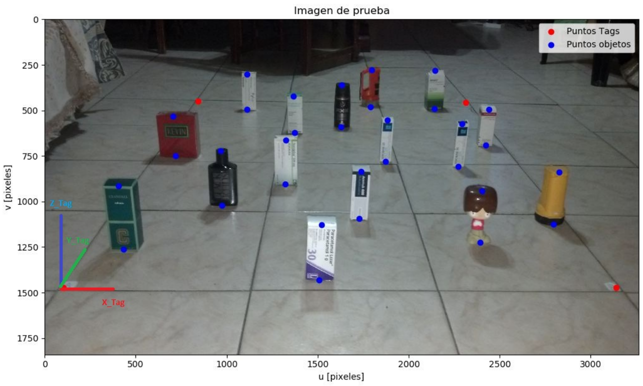


Fig. 2. Test image.

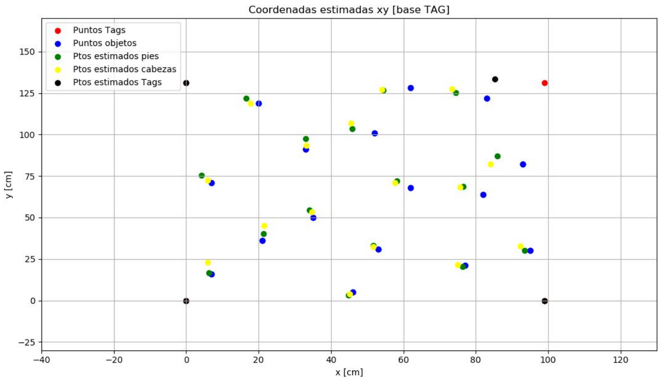


Fig. 3. World xy coordinates. Red dots are the tag points used to generate the ground truth positions of the objects, shown in blue. After optimization our method predicts the positions of the feet (green) and head (yellow) of the pedestrians.

that yield an error of $E_{optim} = 78.53\text{cm}^2$ which means that the projections of feet and head have a standard deviation of ~ 2.2cm. In Fig. 3 we show that the projections of feet and head of the pedestrians in the world's ground plane are reasonably close to each other.

One step left lo mention is that the projections calculated in (7) and (8) are in the world frame of reference whose origin is below the camera, not in the 'tag' frame of reference. The necessary change of base was calculated as an affine transformation [12] such that the tag points projected would coincide with their true world positions.

IV. Conclusion and discussion

We propose a simple approach of calibrating a single view vision system 'in the wild' by just using pedestrian detections. Pedestrians can be detected by many image processing routines and we focus on the calibration that comes after that using the average height of the pedestrians as measuring stick. The calibration parameters allow the back-projections of the feet and head of the pedestrians to the ground and heads plane. The proposed algorithm is to minimise the discrepancy of both back-projections, i.e. the heads and feet of the pedestrians must be vertically aligned.

It is noteworthy that the relationship between the homography that projects from two horizontal planes to the image is as simple as shown in (6).

Clearly the results are encouraging, judging from Fig. 3 the discrepancies between head and feet projections are reasonably small and they both are in general close to the true positions. In general the distance between the estimated back projections and the true positions is similar to the discrepancy between head and feet back-projections, a loose interpretation is that the fit error is similar to the validation error, meaning that the model is constructed reasonably well and that the fit was reasonably good. This toy example shows that the general approach of minimising the discrepancy of head and feet projection is worth developing further.

The roadmap of further development is clear. Mainly the estimation of calibration parameters must be approached from a bayesian perspective, yielding uncertainty estimations instead of the simple point-estimates shown here. This will allow firstly to account for the natural variation of pedestrian heights. Secondly, this will better weight the calibration data, we have seen in [13] the calibration points yield different uncertainties depending on distance to the camera and view factor of the camera with respect to the projecting world plane. A more appropriate weighting of the calibration points based on their associated uncertainties should on average reduce the prediction error.

References

[1] S. Bali and S. Tyagi, “A review of vision-based pedestrian detection techniques,” *International Journal of Advanced Studies of Scientific Research*, vol. 3, no. 9, 2018.

[2] A. Criminisi, I. Reid, and A. Zisserman, “Single view metrology,” *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

[3] M. Evans and J. Ferryman, “Surveillance camera calibration from observations of a pedestrian,” in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, IEEE, aug 2010.

[4] B. Micusik and T. Pajdla, “Simultaneous surveillance camera calibration and foot-head homology estimation from human detections,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, jun 2010.

[5] Z. Zhang, T. Tan, K. Huang, and Y. Wang, “Practical camera calibration from moving objects for traffic scene surveillance,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, pp. 518–533, mar 2013.

[6] I. J. Hales, *Ground plane rectification from crowd motion*. Phd thesis, University of Leeds, 2014.

[7] T. Fäulhammer and P. V. Borges, “A semi-automated framework for homography estimation,” in *Computer Graphics and Imaging / 798: Signal Processing, Pattern Recognition and Applications*, ACTAPRESS, 2013.

[8] D. Forsyth and J. Ponce, *Computer vision: a modern approach*. Always learning, One Lake Street, Upper Saddle River, New Jersey 07458: Pearson Education, 2012.

[9] L. Sobel, “Camera Models and Machine Perception,” tech. rep., Computer Science Department, Technion, 1972.

[10] P. Corke, *Robotics, Vision and Control - Fundamental Algorithms in MATLAB*. Springer, 2011.

[11] T. E. Oliphant, *Guide to NumPy*. Scotts Valley, California, US: CreateSpace Independent Publishing Platform, 2 edition ed., Sept. 2015.

[12] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[13] S. I. Arroyo, U. Bussi, F. Safar, and D. Oliva, “A monocular wide-field vision system for geolocation with uncertainties in urban scenes,” *Engineering Research Express*, vol. 2, p. 025041, jun 2020.