Automatic Extrinsic Calibration of Camera Networks Based on Pedestrians

Anh Minh Truong*
anhminh.truong@ugent.be
TELIN-IPI, Ghent University - imec
Gent, Belgium

Wilfried Philips
wilfried.philips@ugent.be
TELIN-IPI, Ghent University - imec
Gent, Belgium

Junzhi Guan
guanjunzhi@hotmail.com
CETC Key Laboratory of Aerospace Information Applications
Shijiazhuang, China

Nikos Deligiannis
ndeligia@etrovub.be
ETRO, Vrije Universiteit Brussel - imec
Brussel, Belgium

Lusine Abrahamyan
alusine@etrovub.be
ETRO, Vrije Universiteit Brussel - imec
Brussel, Belgium

ABSTRACT

Extrinsic camera calibration is essential for any computer vision tasks in a camera network. Usually, researchers place calibration objects in the scene to calibrate the cameras. However, when installing cameras in the field, this approach can be costly and impractical, especially when recalibration is needed. This paper proposes a novel accurate and fully automatic extrinsic calibration framework for camera networks with partially overlapping views. It is based on the analysis of pedestrian tracks without other calibration objects. Compared to the state of the art, the new method is fully automatic and robust. Our method detects human poses in the camera images and then models walking persons as vertical sticks. We propose a brute-force method to determine the pedestrian correspondences in multiple camera images. This information along with 3D estimated locations of the head and feet of the pedestrians are then used to compute the camera extrinsic matrices. We verified the robustness of the method in different camera setups and for both single pedestrian and multiple walking people. The results show that the proposed method can obtain the triangulation error of a few centimeters. Typically, it requires 40 seconds of collecting data from walking people to reach this accuracy in controlled environments and a few minutes for uncontrolled environments. As well as compute relative extrinsic parameters connecting the coordinate systems of cameras in a pairwise fashion automatically. Our proposed method could perform well in various situations such as multi-person, occlusions, or even at real intersections on the street.

KEYWORDS

extrinsic calibration; camera network; pedestrians;

ACM Reference Format:
Anh Minh Truong, Wilfried Philips, Junzhi Guan, Nikos Deligiannis, and Lusine Abrahamyan. 2019. Automatic Extrinsic Calibration of Camera Networks Based on Pedestrians. In *13th International Conference on Distributed Smart Cameras (ICDSC 2019), September 9–11, 2019, Trento, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3349801.3349802

1 INTRODUCTION

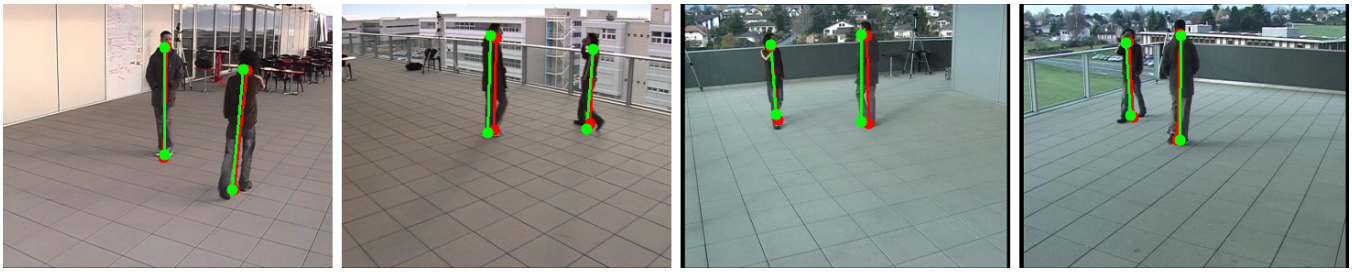
Extrinsic camera calibration provides the coordinate system transformations from 3D world coordinates to 3D camera coordinates for all the cameras in the network. This information is essential for many machine vision applications such as tracking, augmented reality, free view image synthesis, 3D reconstruction [1, 4]. The classical methods [2, 16, 18] require sufficient point correspondences of calibration objects to have accurate extrinsic parameters. Moreover, the objects also have to be well observed from all cameras. Thus, calibrating cameras without mistakes requires a certain level of skill and sending skilled technicians onsite to recalibrate cameras is costly. Their approaches also does not work for historic multicamera video sequences in which no calibration objects were recorded. Hartley et al. [11] proposed an auto-calibration method based on scene reconstruction from arbitrary features. Due to the interactive fashion as well as a large number of parameters to estimate, this method is slow and not always able to achieve reliable results. Many auto-calibration methods [3, 5, 7, 8, 14, 15] based on pedestrians

ICDSC 2019, September 9–11, 2019, Trento, Italy

ACM ISBN 978-1-4503-7189-6/19/09…$15.00
https://doi.org/10.1145/3349801.3349802

(a) Case 1: Camera network 1 - a single person in an empty room

(b) Case 2: Camera network 2 - a single person in an kitchen room



(c) Case 3: EPFL-Terrace dataset [6]

(d) Case 4: Camera network 3 - intersection

Figure 1: Example for detected feet and head positions. Red color represents detected positions by human pose estimation method. Green color represents the detected reprojected positions.

are proposed. However, they are sensitive to noise as well as impractical for many circumstances.

The proposed method relies on finding humans in images and estimating their centerline. For this purpose, we use OpenPose [9, 10], a human pose estimator. Human pose estimation is also an important task for different machine vision applications, such as action recognition, motion capture, sports, etc. Many real-time human estimation methods [9, 10, 17] have been proposed in recent years. Therefore, it does not burden the performance of the system if we combine the human pose estimation with extrinsic calibration for a wide range of applications. The first contribution of our

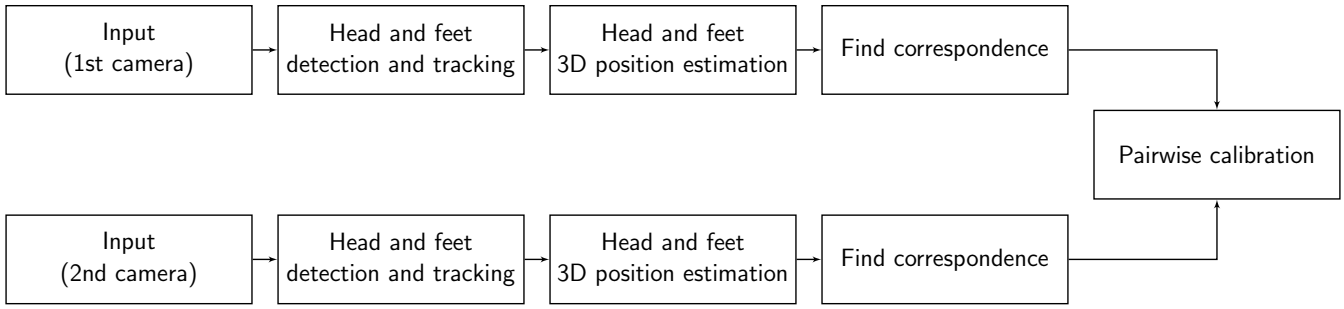


Figure 2: Architecture of the proposed calibration method.

paper is that we replace the ellipse based detection of heads and feet of people by an approach based on modern human pose estimators. This results in a more robust detection of people and a more accurate estimate of their centerlines. We also analyze the amount of time to collect data from walking people to reach the desired accuracy.

A second contribution is that we propose an automatic method to also handle the case of multiple pedestrians simultaneously in the scene. In [7], this case was handled by manual annotation. In the paper, we propose a brute-force, but still fast, method to effectively find the correspondences and also eliminate correspondences which have poor estimated human pose. We show that it produces accurate results and more complicated methods are not needed. A third contribution that we evaluate the proposed calibration method in a wide range of usage scenarios, including indoor and outdoor scenes. The experimental results show that the proposed method could achieve very precise accuracy in both cases.

The rest of the paper is organised as follows. We discuss the related work in Section 1. In Section 2, we describe the architecture of our calibration method for a pair of cameras in detail. In Section 3, we explain the way to extend the proposed method for a camera network. We present the obtained results and the detailed analysis of our experiments in Section 4. Finally, we discuss the conclusion and future work in Section 5.

2 RELATED WORK

Lv et al. [3] detect and select the walking human from video sequences by analyzing the transition of foreground object shapes. They represent the pedestrians as vertical "walking sticks" of the same height in the 3D environment. Then, they compute the vertical vanishing point and the horizon line based on the vertical "walking sticks". Li et al. [8] proposed a single view camera calibration method that directly estimates the focal length, the tilting angle, and the camera height by using a nonlinear regression model from the observed head and feet points of a walking human. In [14], Liu et al. proposed a fully automatic calibration method for monocular stationary cameras. They leverage relative 3D pedestrian height distribution to eliminate false pedestrian detections in moderately crowded scenes. In [15], Liu et al. extended their earlier work to camera network calibration. Iteratively, they incorporate robust matching with a partial direct linear transform. Due to the reliance on vanishing point (intersection of near-parallel lines) estimation, a small error of head (or feet) detection could lead to a big error of extrinsic parameters. On the other hand, our method estimates the extrinsic parameters based on estimated 3D positions of the head and feet which is much less sensitive to noise.

Most methods [3, 14, 15] assume that moving pedestrians walk on a planar, horizontal surface. Possegger et al. [5] proposed an unsupervised extrinsic self-calibration method for a network of static cameras and pan-tilt-zoom cameras solely based on correspondences between tracks of a walking human. Then, they eliminate the outliers of feet and head detection by estimating pairwise homographies between the camera views based on the detected locations of feet and head. Finally, they compute the extrinsic parameters of the cameras by solving a non-linear optimization problem on the reprojection error minimize the reprojection error. Therefore, it tends to get stuck in local optima without a good initialization which was not presented in their work. In contrast, our method could have a precise estimated 3D position of the head and feet based on a robust human pose detector for the extrinsic calibration. Our method also does not requires the person walks on a plane surface. In [13], Lettry et al. proposed a method to solve correspondences for camera calibration based on multiple pedestrian. However, their method could produce incorrect correspondences which degrade the accuracy of the calibration.

Our paper is based on the work of Guan et al. [7]; their method does not require that the pedestrian walks on a plane surface (e.g., walking on steps and stairs), as long as the posture of the pedestrian remains the same while walking. The correspondence of these points between camera views is

assumed to be known. In practice, this method, therefore, requires manual annotations to differentiate between multiple people and is not fully automatic. In this method, head and feet detection is based on change detection and is not very robust w.r.t. noise and occlusion. In contrast, our propose a method is fully automated and uses a more robust human pose detector.

3 PROPOSED METHOD

Note that, we assume that the frame synchronization, as well as the intrinsic calibration for all cameras in the network, have been done before the extrinsic calibration. To obtain the position of a walking person in the image, we apply human pose estimation method [10]. This produces a skeleton model of all major body joints. Because the locations of the head joints are not stable enough in these skeletons, we use the neck joint locations instead as reference positions. We use the midpoint between the left ankle joint and the right ankle joint as the feet position of a walking person.

To obtain the extrinsic parameters for all cameras in the network, we calibrate the camera network in a pairwise fashion. Thus, let us consider a camera system (which is composed of two cameras) where a person moving between N different locations while keeping a fixed posture (the feet and the head of the person can be observed from both cameras). Let $\tilde{\mathbf{u}}_{\mathrm{f}}^{(k)}(t)$ and $\tilde{\mathbf{u}}_{\mathrm{h}}^{(k)}(t)$ be the image positions of the feet and head (neck) at the t-th locations in camera k (where $k \in \{1, 2\}$). Let $\tilde{\mathbf{x}}_{\mathrm{f}}^{(k)}(t)$, and $\tilde{\mathbf{x}}_{\mathrm{h}}^{(k)}(t)$ the normalized image coordinates $(x, y, 1)$ of the feet and the head, respectively. We obtain the unknown Z coordinates of the feet $Z_{\mathrm{f}}^{(k)}(t)$ and Z coordinates of the head $Z_{\mathrm{h}}^{(k)}(t)$ for camera k by applying the proposed method in [7]. Suppose that person walks upright and has heigh h. Let $\mathbf{r}_{\mathrm{h}}^{(k)}(t) = Z_{\mathrm{h}}^{(k)}(t)\tilde{\mathbf{x}}_{\mathrm{h}}^{(k)}(t)$ and $\mathbf{r}_{\mathrm{f}}^{(k)}(t) = Z_{\mathrm{f}}^{(k)}(t)\tilde{\mathbf{x}}_{\mathrm{f}}^{(k)}(t)$ be the 3D camera coordinates of the head and feet. Thus, we have:

$$\mathbf{r}_{\mathrm{h}}^{(k)}(t) - \mathbf{r}_{\mathrm{f}}^{k}(t) = Z_{\mathrm{h}}^{(k)}(t)\tilde{\mathbf{x}}_{\mathrm{h}}^{(k)}(t) - Z_{\mathrm{f}}^{(k)}(t)\tilde{\mathbf{x}}_{\mathrm{f}}^{(k)}(t) = h\mathbf{e}_{z}^{(k)} \quad (1)$$

where $\mathbf{e}_{z}^{(k)}$ is unit vector of the person within camera k. As explained in [7], it is possible to obtain the 3D orientation $\mathbf{e}_{z}^{(k)}$ of the vertical direction (which is the direction parallel to the upright walking persons). From $\tilde{\mathbf{x}}_{\mathrm{f}}^{(k)}(t)$ and $\tilde{\mathbf{x}}_{\mathrm{h}}^{(k)}(t)$, it is possible to compute a 3D vector $\tilde{\mathbf{x}}_{\mathrm{f}}^{(k)}(t) \times \tilde{\mathbf{x}}_{\mathrm{h}}^{(k)}(t)$ which is perpendicular to the unique vertical plane containing the origin of camera k, $\tilde{\mathbf{x}}_{\mathrm{f}}^{(k)}(t)$, and $\tilde{\mathbf{x}}_{\mathrm{h}}^{(k)}(t)$. At a given time instant, the intersection of all of those planes is a line along the vertical direction. Note that, the vectors $\tilde{\mathbf{x}}_{\mathrm{f}}^{(k)}(t) \times \tilde{\mathbf{x}}_{\mathrm{h}}^{(k)}(t)$ are expressed in camera local coordinates, and are therefore transformed with unknown rotation matrices $R^{(k)}$. As shown in [7], these rotation matrices can be recovered using the SVD. Then, the unknown distances $Z_{\mathrm{f}}^{(k)}(t)$ and $Z_{\mathrm{h}}^{(k)}(t)$ by solving a set of equations; from this the camera translation matrices can be computed. A problem with any extrinsic calibration technique not using calibration objects is that it produces coordinate transforms which are defined up to an unknown scale factor only. However, we can then scale all results with a standard height to obtain extrinsic parameters that match the world coordinate.

Algorithm 1: Compute matching rate of the extrinsic parameters between camera a and camera b

```
total number of pairs n_pairs ← 0 ;
total number of matched pairs n_matched ← 0 ;
forall time steps t do
    H^(a)(t) ← pairs of head and feet locations of
      camera a at time step t ;
    H^(b)(t) ← pairs of head and feet locations of
      camera b at time step t ;
    C^(ab)(t) ← combinations of H^(a)(t) and H^(b)(t) ;
    forall combination c in C^(ab)(t) do
        matched(c) ← 0 ;
        forall pair p of head and feet locations in c do
            if reprojection error of p < threshold then
                matched(c) increased by 1 ;
    length_a ← number of pairs in H^(a)(t) ;
    length_b ← number of pairs in H^(b)(t) ;
    n_matched increased by max_c(matched(c)) ;
    n_pairs increased by min(length_a, length_b);
return matching rate ← n_matched/n_pairs ;
```

We use Openpose [10] to estimate 2D skeleton models of humans in the images. In practice, the estimated locations of the necks and feet are inconsistent between views (e.g., a different physical point is indicated for a foot in two views). Thus, the calibration results would be poor if people are observed in an insufficient number of locations (e.g., the method will fail if only a single person, always in the same position, is observed). In controlled environments, these conditions can be easily enforced by providing instructions to the walking people. In uncontrolled environments, many people tend to pass the scene in a short amount of time such as walking along a straight line (insufficient number of locations). Hence, it is difficult to gather data in different locations of the scene for a precise calibration.

To handle the case of multiple people being present simultaneously. We propose an easy and robust method to

Table 1: Comparison between our method, the method of Guan et al. [7], and the method of Hödlmoser et al. [12]. We randomly select 20 locations of the pedestrians in the scene to calibrate the CN1, CN2. For CN3 and EPFL-Terrace (CN4), we apply the proposed method on the first half of the video. $\delta\mathbf{r}^{(w)}$, $\delta\mathbf{u}^{(p)}$, $\delta\mathbf{u}^{(r)}$, and $\delta\mathbf{u}^{(rr)}$ denotes the triangulation error, projection error, reprojection error, and relative reprojection error, respectively.

	Proposed method				Guan et al [7]				Hödlmoser et al. [12]			
	CN1	CN2	CN3	CN4	CN1	CN2	CN3	CN4	CN1	CN2	CN3	CN4
$\delta\mathbf{r}^{(w)}$ (cm)	1.33	**2.2**	-	-	**1.30**	3.16	-	-	**1.30**	63.7	-	-
$\delta\mathbf{u}^{(p)}$ (pixel)	**3.98**	**5.8**	-	-	4.14	6.72	-	-	4.15	106.4	-	-
$\delta\mathbf{u}^{(r)}$ (pixel)	**3.76**	**5.0**	-	-	4.09	6.20	-	-	4.09	104.3	-	-
$\delta\mathbf{u}^{(rr)}$ - head (%)	**1.8**	**1.7**	12.6	2.1	1.9	7.0	-	-	1.9	43	-	-
$\delta\mathbf{u}^{(rr)}$ - feet (%)	**2.8**	**2.0**	17.2	2.3	3.0	4.1	-	-	3.0	39	-	-

solve the association problem (which pedestrian in one camera corresponds to an observation in another camera). First, we apply a simple object matching algorithm based on feature matching to track the pedestrians for each camera. Let $H_i^{(k)} = \{(\tilde{\mathbf{u}}_{\mathrm{f}}^{(k)}(m), \tilde{\mathbf{u}}_{\mathrm{h}}^{(k)}(m)) \cdots, (\tilde{\mathbf{u}}_{\mathrm{f}}^{(k)}(n), \tilde{\mathbf{u}}_{\mathrm{h}}^{(k)}(n))\}$ be the set of all locations of the head and feet of person i from frame m to frame n in camera k with $k \in \{a, b\}$. Furthermore, let $H^{(k)} = \{H_0^{(k)}, H_1^{(k)}, \cdots, H_q^{(k)}\}$ be a set of locations head and feet of all pedestrians in the scene of camera k with q is the number of pedestrians in this scene. We compute all possible correspondences $C^{(ab)}$ between camera a and camera b by generating all pairs of elements from $H^{(a)}$ and $H^{(b)}$. Then, we calibrate the pair of cameras with each generated correspondences. The matching rate of the extrinsic parameters is defined in Algorithm 1. Finally, the top highest matching rate correspondences are selected to calibrate the pair of cameras. The proposed method not only find correct correspondences but also remove correspondences which have poor human pose estimation. Therefore, it also improves the robustness of extrinsic calibration. To deal with the combination explosion, the frames with too many pedestrians (where the pose estimator also have a poor result) are removed. In practice, we only need to calibrate the network once before processing the data from the network. Hence, we could calibrate the camera network from the scene which has low number of pedestrians. Therefore, the number of combination for each frame is always less than 120 which does not take too much time to verify all possibilities. Thus, it does not limit the scalability of the proposed method.

4 EXPERIMENTAL RESULTS

Performance Measures: In order to evaluate the performance of our method with ground truth points, we compute the triangulation error ($\delta\mathbf{r}^{(w)}$), projection error ($\delta\mathbf{u}^{(p)}$), and reprojection error ($\delta\mathbf{u}^{(r)}$) [7]. In practice, the ground truth points are not always available to measure the performance of the calibration. Thus, we can only measure the calibration by computing reprojection error based on the head (or feet) positions of detected pedestrians. However, different cameras have different resolution. Moreover, the height of the pedestrians at different locations in an image is different. Hence, to deal with this problem, we define the relative reprojection error as follows:

$$\delta\mathbf{u}^{(rr)} = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{\left\| \mathbf{u}_{mk} - \hat{\mathbf{u}}_{mk}^{(rr)} \right\|_2}{h_{mk}} \tag{2}$$

where N be the number of ground truth 3D sample points, K is the number of cameras in the network, M is the number of pedestrians, h_m is the height in the image of person m-th in the camera k-th. $\mathbf{u}_{mk}$ is the observed pixel coordinates of the head (or feet) of person m-th in the camera k-th. $\hat{\mathbf{u}}_{mk}^{(rr)}$ is the estimated location of head (or feet), which are obtained through reprojection.

Calibration with single person: We evaluate our method with a multi-camera tracking system composed of four side view cameras. For simplicity, we call it Camera Network 1 (CN1). The cameras were mounted at a height of about 3 m at each corner of a room (8.6 m by 4.8 m). The resolution of the videos are 780 by 580 pixel (Figure 1a). We obtain the intrinsic parameters by [18]. In this paper, we compare our method to the calibration method of Hödlmoser et al. [12] and Guan et al. [7]. Figure 1 shows an example of the detected head and feet positions of the person in a scene. Table 1 shows that our method has more accurate results than the state-of-the-art methods. We also show that the proposed method requires the person to walk around the room for 2-3 times to achieve a stable and accurate calibration. (Table 2). For single person case, we apply the refinement method which proposed in [7] to obtain the final extrinsic parameters (Table 1).

In order to show that our method work in a complex real-life environments room setup, we also evaluate our method on a three camera setup in a kitchen (Figure 1b). The cameras

were mounted at a height of about 2 m at different corners of a room. The resolution of the videos are 640 by 480 pixel. We call it Camera Network 2 (CN2) for simplicity. Table 1 shows that our method outperforms the method proposed by Guan et al. [7]. Note that, the person in this scene was cleaning the kitchen floor which is a realistic circumstance. It shows stability, robustness of our method to the occlusion.

Calibration with multiple pedestrians

Table 2: Success percentages (the triangulation error is below 15 cm) within 1000 experiments of the proposed method for the CN1.

Moved distance (m)	6.5	10	20	25	40
Successful percentage (%)	35	83	97	99	100

We evaluate our calibration method on the [6], which is a public multi-camera pedestrians video dataset (Figure 1c). This dataset includes two sequences, which were shot outside our building on a terrace with four DV cameras. To show that our method can be applied to a real-life situation, we also record several video sequences at an intersection in Ghent to evaluate the proposed method (Figure 1a). The pedestrians in this scene are quite small (about 60 pixels height). We call it Camera Network 3 (CN3) for simplicity.

Table 1 shows our method has reasonably low error among different circumstances. However, in the intersection case, the pedestrians appear in some regions are too small to detect by the human pose estimation. In addition, when the trajectories of the pedestrians are too short, the matching rate of them are too small to be selected. Hence, the proposed method also could not match the correspondence between the cameras which leads to the higher relative reprojection error at some regions of the scene. However, the errors in these regions are still acceptable for the applications of this type of scenes. It only takes approximately 270 seconds and 210 seconds on EPFL-Terrace dataset and CN3 to solve the correspondences and obtain the extrinsic parameters, respectively (we implemented the code to run on CPU with Python).

5 CONCLUSION

In this paper, we present a simple and robust method to leverage the human pose estimation for the 3D positions of the head and feet computation. To handle the case where multiple pedestrians are in the scene, we also developed a brute-force method to select appropriate head and feet locations for the extrinsic camera calibration. The proposed method could work well in various environments and is robust against occlusion compared to state-of-the-art methods. More importantly, the proposed method could work completely automatically without manually selecting proper input data for the calibration method. In the future, we will investigate a regional selection method to handle the case where the walking trajectory is too short.

ACKNOWLEDGMENTS

This work was financially supported by the Flemisch Fund for Scientific Research FWO-Flanders through the grant 3G014718.

REFERENCES

[1] Martin Dimitrievski, Peter Veelaert, and Wilfried Philips. 2019. Behavioral pedestrian tracking using a camera and lidar sensors on a moving vehicle. *SENSORS* 19, 2 (2019), 34.

[2] Ernest L. Hall et al. 1982. Measuring Curved Surfaces for Robot Vision. *Computer* 15, 12 (1982), 42–54.

[3] Fengjun Lv et al. 2006. Camera calibration from video of a walking human. *IEEE Trans. Pattern Anal. Machine Intell.* 28, 9 (2006), 1513–1518.

[4] Gaurav Chaurasia et al. 2013. Depth Synthesis and Local Warps for Plausible Image-based Navigation. *ACM Trans. Graph.* 32, 3 (2013), 30:1–30:12.

[5] Horst Possegger et al. 2012. Unsupervised Calibration of Camera Networks and Virtual PTZ Cameras. In *Proceedings of the Computer Vision Winter Workshop*.

[6] Jerome Berclaz et al. 2011. Multiple Object Tracking Using K-Shortest Paths Optimization. *IEEE Trans. Pattern Anal. Machine Intell.* 33, 9 (2011), 1806–1819.

[7] Junzhi Guan et al. 2016. Extrinsic calibration of camera networks based on pedestrians. *SENSORS* 16, 5 (2016).

[8] Shengzhe Li et al. 2015. A simplified nonlinear regression method for human height estimation in video surveillance. *EURASIP Journal on Image and Video Processing* 2015, 1 (2015), 32.

[9] Zhe Cao et al. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[10] Zhe Cao et al. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[11] Richard I. Hartley. 1994. An algorithm for self calibration from several views. In *CVPR*. 908–912.

[12] Michael Hödlmoser and Martin Kampel. 2010. Multiple Camera Self-calibration and 3D Reconstruction Using Pedestrians. In *Advances in Visual Computing*. Springer, Berlin, Heidelberg, 1–10.

[13] Louis Lettry, Ralf Dragon, and Luc Van Gool. 2016. Markov chain Monte Carlo cascade for camera network calibration based on unconstrained pedestrian tracklets. *Proceedings ACCV 2016* 10112, 1–15.

[14] Jingchen Liu, Robert Collins, and Yanxi Liu. 2011. Surveillance camera autocalibration based on pedestrian height distributions. In *Proceedings of the British Machine Vision Conference*.

[15] Jingchen Liu, Robert Collins, and Yanxi Liu. 2013. Robust autocalibration for a surveillance camera network. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*. 433–440.

[16] R. Tsai. 1987. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal on Robotics and Automation* 3, 4 (1987), 323–344.

[17] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple Baselines for Human Pose Estimation and Tracking. In *ECCV*.

[18] Zhengyou Zhang. 2000. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Machine Intell.* 22, 11 (2000), 1330–1334.