2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance

Surveillance Camera Calibration from Observations of a Pedestrian

Murray Evans and James Ferryman
Computational Vision Group
School of Systems Engineering
University of Reading
Reading, UK
{m.evans@reading.ac.uk | j.m.ferryman@reading.ac.uk}

Abstract

Calibrated cameras are an extremely useful resource for computer vision scenarios. Typically, cameras are calibrated through calibration targets, measurements of the observed scene, or self-calibrated through features matched between cameras with overlapping fields of view. This paper considers an approach to camera calibration based on observations of a pedestrian and compares the resulting calibration to a commonly used approach requiring that measurements be made of the scene.

1. Introduction

Camera calibration is a useful task for many computer vision applications, permitting real-world measurements of the scene to be made, and more importantly, providing a robust relationship between the multiple cameras that may exist and the scene being observed. There are two parts to any calibration process. The first is to determine the intrinsic parameters of the camera. These parameters control the model of the imaging process, or how a point in space defined relative to the camera projects to the image. The second part of the calibration process is to determine the extrinsic parameters of the camera, or the pose. These parameters detail the location and orientation of the camera in space relative to the rest of the environment.

Generally, cameras can be represented by the pinhole model as described in many textbooks [8], often augmented by various distortion parameters to account for any warping introduced by the camera lens. The intrinsic parameters of this model are often estimated using one of several available toolkits and a calibration target [2]. Often these approaches will require access to the camera (being able to get close enough that the calibration target is large enough in the resulting image), and become impractical in many surveillance situations when the cameras are already installed. They also focus more on providing the intrinsic calibration of the camera, requiring further efforts to determine the extrinsic parameters.

Surveillance cameras can be robustly calibrated both intrinsically and extrinsically using the Tsai method [16], but this requires, at the very least, measuring the position of significant features of the scene visible to the camera(s), and may require augmenting the scene with suitable features as well. This takes time and effort, which may need to be repeated should the camera be moved due to environmental effects or maintenance. It may also be less than simple to make measurements of points in the environment, particularly when the ground is non-planar.

Cameras can also be *self-calibrated* [10] when one or more viewpoints is available. In this case, features of the scene are matched between views, and this leads through a chain of processing to the calibration of the camera(s). Such a method is quite reliable for single cameras that move through the scene, but the automatic matching of features in wide baselines is difficult even with robust features such as SIFT [13]. A further complication is that a key stage of the approach is the estimation of the Fundamental Matrix which requires that corresponding points in images must lie on more than one scene plane. Surveillance scenes can often be relatively large, flat, open spaces best described by a single plane, meaning the scene may need to be augmented by the presence of objects of interest before suitable matching features can be acquired.

As such, there has been much interest in calibrating cameras from information gleaned from tracked pedestrians (a typical object to be added to a surveillance scene), where [3, 11, 14] are typical examples. This paper will discuss how a useful calibration can be determined from an observed pedestrian for a scene with multiple cameras and overlapping fields of view. The process can be summarised as 1) gather useful information from observing the pedestrian, 2) determine the intrinsic parameters of each camera, 3) determine the relative extrinsic parameters for each pair

978-0-7695-4264-5/10 $26.00 © 2010 IEEE
DOI 10.1109/AVSS.2010.32

of cameras, 4) combine the cameras into a single camera network and optimise the calibration parameters through a bundle adjustment.

2. The Pedestrian

The pedestrian is, potentially, a very useful tool for calibrating cameras. Firstly, they are a very flexible and easily obtained calibration target. Secondly, they are generally of good scale in any surveillance type scenario. Finally, they more or less represent a vertical line in space, and it is this vertical line that first drew attention as being useful for calibration.

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. When walking, the point on the ground becomes much more ambiguous and the person typically tends to lean forwards.

One simple method for determining the desired head and foot positions is to use a standard background subtraction method, fitting an ellipse to the blob from the required person, and determining where the major axis of this ellipse intersects the smallest bounding box that encompasses the whole of the pedestrian. Fully automating this process becomes difficult when the scene can not be emptied of all but the target person (it is not simple to automatically determine which person should be observed across numerous images with no calibration to relate the cameras), however, even in scenes that are extremely busy it is trivial, if tedious, to manually acquire a suitable set of points. Figure 1 shows an example of a set of points extracted for the PETS2009 [6] dataset. Producing this set of points for four cameras took less than half an hour, whereas it is reported to have taken several hours to make the measurements for the Tsai calibration supplied in the dataset.

3. Intrinsic Calibration

There exists significant published work on the intrinsic calibration of cameras. With respect to observations of a pedestrian, the significant approach is typically to make use of vanishing points and vanishing lines [3, 11, 14]. Firstly, the person is assumed to be a vertical line, and secondly, it is assumed that the person is walking on a flat ground plane, the normal of which is parallel to the person's vertical line.

If the intrinsic parameters of the camera are restricted to the focal length and principal point, making the reasonable and common assumption that the camera's pixel sensors are square and introduce no skew, then it is known that they can be estimated from three vanishing points, where each vanishing point is in an orthogonal direction to each other [8, 9, 14]. If there are two orthogonal lines on the

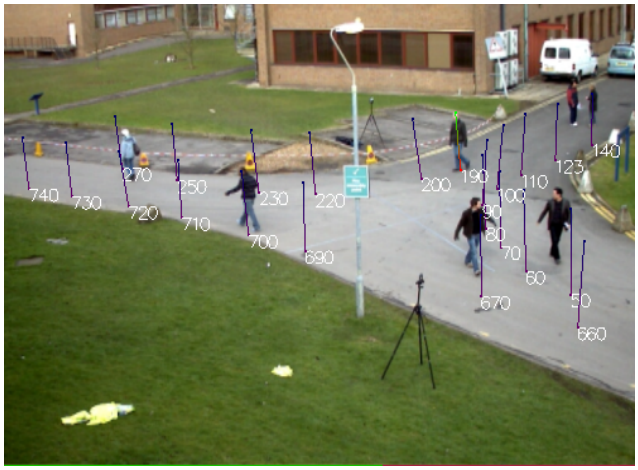

Figure 1. Extracted head and foot points for a pedestrian in the PETS2009 dataset.

ground plane which can be extended to infinity and intersected with the horizon visible in the image, then that provides two of the vanishing points. The third comes from observations of the person. Alternatively, if the principal point is assumed to be the centre of the image then the focal length can be determined from a vanishing point and a vanishing line [11], while with two views, the vanishing line - vanishing point combination can provide focal length and principal point [3].

Both the vanishing line and vanishing point are, in theory, easily computed from the pedestrian, as shown in Figure 2.

The vanishing line can be robustly calculated. Any two positions for the person produces two (head point, foot point) pairs $(\mathbf{h}_1, \mathbf{f}_1)$, $(\mathbf{h}_2, \mathbf{f}_2)$. A line l_h defined by the two head points, and a line l_f defined by the two foot points will intersect, and if the person walks on a flat ground plane, this intersection point will lie on the vanishing line for the ground plane. Given that the pedestrian can be observed in many locations, a large number of vanishing points can be found. A line fit to these vanishing points is a robust estimate of the vanishing line of the ground plane.

The process for estimating a vertical vanishing point for lines orthogonal to the ground plane is in concept similarly simple. At any point on the ground plane, the line through the person's head and foot extends to infinity. Given sufficient observations of the person the lines should intersect at the appropriate vanishing point. This process however is extremely sensitive.

To begin with, a person is never really vertical (the extracted head and foot points visible in Figures 1 and 3 will testify to this), and to further complicate matters, the vanishing point can very often be a very long distance from the image centre. Indeed, if the camera is low enough, all the head/foot lines could be vertical in the image, and the van-

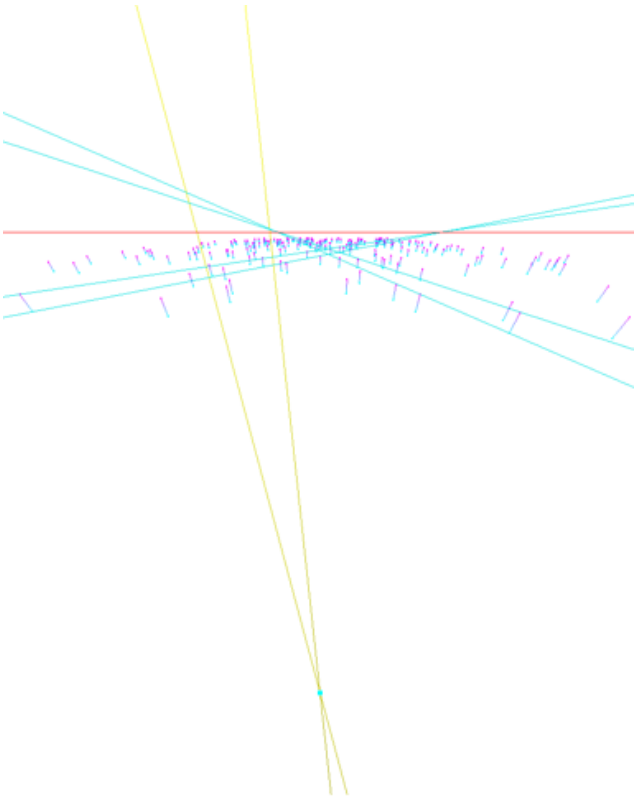

Figure 2. Short lines indicate lines between synthetically generated head and foot positions. Cyan lines crossing the image in pairs on diagonals are lines between two pairs of person positions. One line of each pair connects the heads, one the feet. The intersection of the head-head foot-foot lines gives a vanishing point. The two vanishing points define the horizontal red vanishing line. Long, near vertical yellow lines are an extention of the head-foot lines of two positions. These intersect at a vertical vanishing point.

shing point approaches infinity. This means that any slight inaccuracy in the position of a head or foot point by even as much as half a pixel can lead to large offsets in the estimate of the vanishing point. While [11] seeks to improve this accuracy through more robust tracking of the pedestrian, one still has to consider how lens distortions which have yet to be considered would affect the result. Thus it is perhaps simpler and more reliable to use an alternative approach.

If an assumption is made about the principal point of two cameras, then a simple equation can lead to the focal lengths when a Fundamental Matrix is known between the cameras [1]. This too however can be problematic [8] and can produce unusable values in practice.

As such, this paper proposes, for the sake of practicality, that the intrinsic parameters of the cameras be completely set to some assumed value unless they can be internally calibrated off line prior to installation in the scene. Experience then suggests that a useful calibration can be achieved by relying on the bundle adjustment to determine an internal calibration. The further advantage to this is that one is no longer constrained to require the pedestrian to move only on a single plane, and the head and ground point need no longer actually be vertically aligned – they just have to be correctly corresponding between views.

4. Relative Extrinsic Calibration

The extrinsic parameters relating two cameras can be determined through use of the epipolar geometry. Estimating the Fundamental Matrix between two views depends on knowing a set of point correspondences between the views. These can be supplied by the observations of the pedestrian, with each observation supplying two corresponding points between the views - the head point and the foot point. In conjunction with a Robust RANSAC process [7], the 8-Point algorithm [8] can be used to reliably determine the Fundamental Matrix F for the two cameras.

Ultimately, what is desired is the translation $\mathbf{t}$ and rotation R that will take the first camera to the position of the second camera. These are not directly available from the Fundamental Matrix, but if the intrinsic parameters for each camera are known then F can be decomposed to the Essential Matrix E, and E can be further decomposed to R and $\mathbf{t}$.

4.1. Decomposing the Fundamental Matrix

Firstly, define the *camera matrices* K_1 and K_2 for the two cameras

$$K_1 = \begin{bmatrix} f_1 & 0 & p_{x1} \\ 0 & f_1 & p_{y1} \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

$$K_2 = \begin{bmatrix} f_2 & 0 & p_{x2} \\ 0 & f_2 & p_{y2} \\ 0 & 0 & 2 \end{bmatrix} \tag{2}$$

where f_1 and f_2 are the focal lengths of each camera, and (p_{x1}, p_{y1}) and (p_{x2}, p_{y2}) are the principal points of the two cameras. The Essential Matrix can then be computed from the Fundamental Matrix F as:

$$E = K_2^T F K_1 \tag{3}$$

4.2. Decomposing the Essential Matrix

Hartley [8] details the decomposition of the Essential Matrix using the singular value decomposition,

$$E = U \begin{bmatrix} d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & 0 \end{bmatrix} V^T \tag{4}$$

The translation can be recovered as $\mathbf{t} = \pm \mathbf{v}_3$, where $\mathbf{v}_3$ is the final column of V. Note the unknown sign of $\mathbf{t}$.

The rotation is a 3×3 matrix R, which can be either:

$$R_1 = UWV^T \tag{5}$$

or,

$$R_2 = UW^TV^T \tag{6}$$

where,

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{7}$$

It is not often noted in the literature that this decomposition of E can lead to improper rotations (a rotation matrix with negative determinant). Such a decomposition is inherently wrong. Should either of the possible rotations be improper, a correct solution can be obtained be negating the matrix E and repeating the decomposition.

This results in four possible solutions for the pose of the second camera: $(+t, R_1)$, $(+t, R_2)$, $(-t, R_1)$, $(-t, R_2)$. Only one of these solutions will cause all of the available point correspondences to project to a position in front of both cameras. While only one correspondence should be enough to determine the correct solution, the failure to reproject all points in any solution can form a good indicator that the decomposition has failed to produce a good solution. A total failure in this way generally indicates that the intrinsic parameters, which up to this point have only been guessed at, were incorrect. If the decomposition does not fail, there is still a question of just how accurate the resulting translation and rotation can be given the assumed internal parameters. Experience suggests that if the recovered translation is generally in the correct direction, the remainder of the calibration process will be successful. Typically, this will be the case when the decomposition is successful and all the points can be reconstructed in front of both cameras.

5. Extrinsic Calibration

At this stage, there exists a set of pairs of cameras for which R and t were successfully estimated. A further assumption has been made that the distance between the cameras of each pair is one unit. Though ultimately this will not be true for all the camera pairs, at this stage, this is not estimated, and the remaining stages will correct the distances between cameras.

Take one camera as the reference camera, and set its position to be the origin $\mathbf{c}_r = [0, 0, 0]^T$ and to have rotation the identity matrix $R_r = I$. This could be the camera with the largest number of successfully paired cameras, or the one with the consistently smallest error when estimating the Fundamental Matrix. This forms the first *positioned* camera. Cameras can be iteratively added to the *positioned* set if they are part of a pair with one other camera already positioned. If $\mathbf{c}_p$ and R_p is the position and orientation of the camera already in the positioned set, and $\mathbf{t}$ and R the relative translation and orientation of the new camera from the positioned camera, then the new camera's position can be set as $\mathbf{c}_n = \mathbf{c}_p + R_p\mathbf{t}$ and orientation as $R_n = R_pR$.

With each of the camera poses initialised the aim is now to optimise this network of cameras such that the camera calibrations and image point data are consistent and accurate. To achieve this, a bundle adjustment [15] is used. However, just throwing all the cameras and all the available points at the optimisation in one go will generally lead to failure. A superior approach is to increase the complexity of the optimisation in stages.

Firstly, select the reference camera and one other camera. These two cameras will share a certain portion of the head/foot points, and the initial pose estimate of the two cameras can be used to reconstruct these head/foot points in 3D. A first run of bundle adjustment can now be performed to optimise the parameters of the first two cameras as well as the 3D points. Once these cameras have been optimised one can add a third camera, then new points observable from the third camera, then a fourth camera and so on until finally all the cameras have been added. Most success has been had when, out of all the available intrinsic parameters only the focal length is permitted to change in the initial optimisations. Only once all of the cameras have been added are the principal point and distortion parameters added to the optimisation. A variety of bundle adjustment implementations are available for use, the most widely known being [12], though results for this paper were generated using [17].

5.1. Setting the Scale

The reconstructed 3D points for the pedestrian provide a mechanism to determine the as yet unknown scale in the calibration. The distance between the head and foot point for any one position of the person should be approximately the actual height of the person. If the person is available, then their height can be measured, if not, a reasonable estimate of the average height of a person could be used instead, albeit with the caveat that any measurements made using the calibration will be innaccurate if the observed pedestrian was abnormally tall or short. The distance between cameras should be scaled such that the average distance between head/foot pairs equals the chosen height.

Assuming that reasonable care is taken in creating the

head/foot positions for each image (making sure that corresponding image points are actually images of the same scene point, and that the cameras were reasonably well synchronised) this approximation of the scale factor for the scene is the largest source of error, at least in terms of making accurate length measurements of the scene. In many surveillance scenarios, making an accurate measurement of a length is of less importance than correctly relating data between viewpoints and fusing that data; knowing that a distance is approximately one metre is typically just as useful as knowing that it is precisely one metre. As such, it can be said that this pedestrian based approach is useful and practical in that it fulfills the need to reliably relate data between the cameras. If more precise measurement is needed, one could make a few precise measurements of the scene for the sole purpose of setting the scale.

5.2. Transforming to Preferred Coordinate System

Generally speaking, it is not desirable to describe the environment in terms of the camera's coordinate system. Rather, the cameras should be described relative to some scene origin. Often the observed scene is approximated by a plane, and it is desirable to have the scene origin on that ground plane. If the ground is assumed to be planar, then each foot point reconstructed during the bundle-adjustment process should form a plane in space. The aim is to transform this plane to be the ground plane $z = 0$, in the processes transforming the whole camera system to be relative to the scene origin defined on this plane.

Assuming the scene origin is visible in the view of one of the cameras, transforming the camera network from being relative to the initial position of the reference camera to being aligned with the scene coordinate system is fairly simple. Let $\hat{\mathbf{o}}$ be the image of the scene origin as seen by camera c. To ensure the proper orientation of the final camera network an image of a point on the scene y or x axis also needs to be known. Let $\hat{\mathbf{y}}$ be the image of a point known to lie on the scene y axis (that is some point $[0, y, 0]^T$ in space, for any scalar value y).

The first step is to estimate the ground plane in the initial camera coordinate system. This is done by simple least squares fitting to the reconstructed foot points. The calibration of camera c can then be used to back-project the image of the scene origin $\hat{\mathbf{o}}$ to the foot plane, providing the location $\mathbf{o}$ of the scene origin relative to the initial camera coordinate system.

The second step is to translate the camera system such that $\mathbf{o}$ becomes the point $[0, 0, 0]^T$, which is done by applying the translation $\mathbf{t} = -\mathbf{o}$ to each camera.

Next, the aim is to ensure that the foot plane becomes the plane $z = 0$, which means finding a rotation that will transform the foot plane normal $\mathbf{n} = [a, b, c]$ to be $[0, 0, 1]^T$. If $||\mathbf{n}|| == 1.0$, then the angle of the rotation can be determined from the dot product $cos(\theta) = \mathbf{n} \cdot [0, 0, 1]^T$, while the axis of the rotation is given by the cross product $\boldsymbol{\alpha} = \mathbf{n} \times [0, 0, 1]^T$. Standard methods exist to convert this angle-axis representation into a standard rotation matrix R. If $\mathbf{c}$ is the position of a camera in space, then the position of the camera should be updated to $\mathbf{c} \leftarrow R\mathbf{c}$. Similarly, the orientation of the camera R_c should be updated to $R_c \leftarrow RR_c$. Once transformed, all cameras will have a pose such that the foot plane is the plane $z = 0$.

Finally, one needs to ensure that the orientation of the camera system matches that of the desired scene coordinate system. This involves ensuring that the normalised direction from the origin to any point on the y-axis is the direction $[0, 1, 0]^T$. To determine this, back project the image point $\hat{\mathbf{y}}$ to the foot plane, to determine a point $\mathbf{y}$. If all previous rotations and translations were correct, then $\mathbf{y}$ should be some point $[x, y, 0]^T$. Let $\mathbf{d} = \mathbf{y} - \mathbf{o}$, then let $\mathbf{d}_n = |\mathbf{d}|$ be the normalised direction. Again, the required angle of the rotation comes from the dot product $cos(\theta) = \mathbf{d} \cdot [0, 1, 0]^T$, while the axis of the rotation comes from the cross product $\boldsymbol{\alpha} = \mathbf{d} \times [0, 1, 0]^T$. Applying this rotation to the camera positions and orientations should now ensure that the camera system is aligned with the desired scene coordinate system.

6. Evaluation

The practicality of this approach is only valid if the resulting calibration is reliable enough for use. As such, the approach can be compared to results from a Tsai calibration of the same scene. Suitable scenarios are provided by the PETS 2006 [4], 2007 [5] and 2009 [6] data sets, as well as an airport data set recorded at Toulouse. Useful calibrations have also been achieved in similar situations where no other calibration was available for comparison. In each case, the intrinsic parameters for the cameras were initialised as:

$$K = \begin{bmatrix} 500 & 0 & w/2 \\ 0 & 500 & h/2 \\ 0 & 0 & 1 \end{bmatrix} \quad (8)$$

where w is the width of the image in question, and h the height.

6.1. Airport Data Set

This data set consists of eight static cameras installed around an apron at Toulouse Airport. The cameras have varying degrees of overlap with each other, with one camera suffering significant occlusion of a large portion of the scene due to presence of the jet-bridge. In this case the pedestrian walking the scene was explicitly tasked for the calibration process and traverses a route that covers as much of the scene as possible. Figure 3 shows most of the path traversed from the point of view of one of the cameras that sees most of the scene. Figure 4 shows the reconstruction

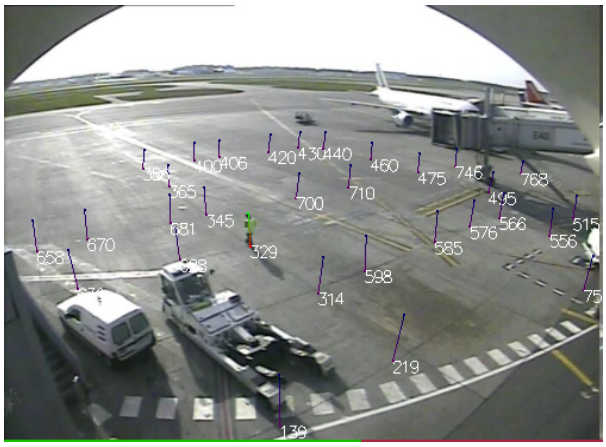

Figure 3. Path walked by the pedestrian in the Airport data set.

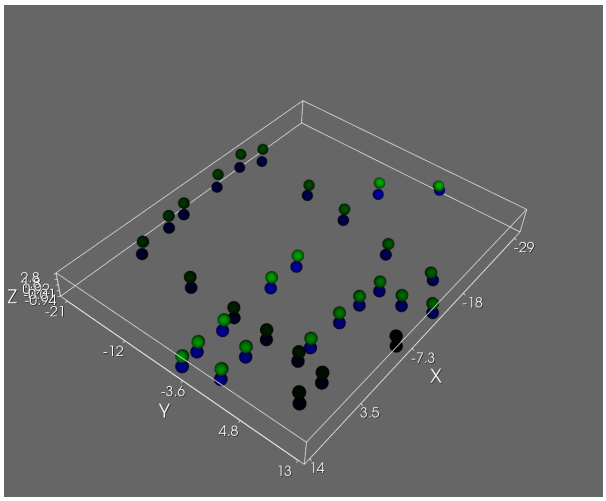

Figure 4. 3D reconstruction of the head and foot points of the pedestrian observed in the Airport data set.

of the head/foot points. As can be seen, these correctly fall into two planes - one for the heads and one for the feet, with each pair of head/foot points appearing very closely aligned vertically.

To get some estimate of the error of the calibration, the following approach is used. For a single view, each visible foot point is back-projected using either the Tsai or Pedestrian based calibration to the foot plane. The resulting 3D points are then projected into each other view. Because the correct corresponding image point is known in each view, an error can be computed as the difference between the observed image point and the projected 3D point. Thus the mean error between any two views can be determined. Tables 1 and 2 show the results of this test on the airport scene (for space reasons, only the mean against all cameras is shown rather than the error against each individual camera). In this case the calibration based on pedestrian observations is significantly more consistent across the views

camera	1	2	3	4
mean	6.064	2.329	3.274	3.519

camera	5	6	7	8
mean	3.176	3.958	3.946	3.781

Table 1. Mean error, in pixels, for each camera against all the other cameras, for the pedestrian calibration for the Airport dataset.

camera	1	2	3	4
mean	n/a	6.315	10.844	6.333

camera	5	6	7	8
mean	5.602	26.006	22.043	4.115

Table 2. Mean error, in pixels, for each camera against all the other cameras, for the Tsai calibration (calibration not available for camera 1) using the Airport dataset

than that of the Tsai calibration. This is perhaps not too surprising, firstly because the Tsai calibration is performed on each view independently, and secondly, because a number of the manually set image points were inaccurately positioned (they did not correctly correspond to the scene point for which a measurement was known, in some cases being as much as ten pixels in error).

To test how well the scale estimate works, and thus how closely a distance measurement made using the calibration matches up to a real distance, a further test is undertaken. It was firstly ensured that the axes and origins of the pedestrian based calibration and the Tsai based calibration were correctly aligned. The measured calibration points used for creating the Tsai calibration provide actual measurements of a set of points in the scene. This consists of a set of ground points in the scene with positions measured relative to the scene origin, and the correlating image points in each view. The distance of each measured scene point from the origin is thus known. For each view, the pedestrian based calibration can be used to back-project the image points onto the ground plane of the scene. The distance of each of these back projected ground points from the origin can thus be deduced. Allowing for the fact that some of the image points were inexpertly placed in the Tsai calibration data, it was found that the scale of the scene as measured by the pedestrian calibration was consistently about 0.95 the scale of the measured scene. The important point here is the consistency, as this suggests that the difference in scale arrises primarily from a slight error in the estimate of the person's height, and indeed, adjusting the height estimate saw the calibration fall closer in line with the scale of the actual scene.

6.2. PETS 2009

This data set consists of four permanently installed CCTV cameras, as well as four camcorders on tripods. Taking just the four CCTV cameras, Figure 1 has already shown

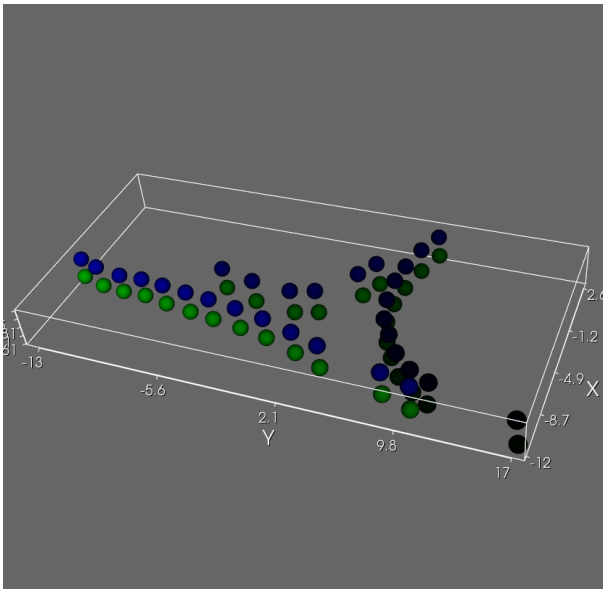

Figure 5. 3D reconstruction of the head and foot points of the pedestrian observed in the PETS 2009 data set.

the path taken by the observed pedestrian. Figure 5 shows the reconstruction of the head and foot points in 3D after the calibration process. Again, the points correctly form two planes, one for the heads, and one for the feet, and closely follow the path walked by the pedestrian.

Tables 3 and 4 show the error between each view for both calibrations. As with the airport scene, the pedestrian based calibration is comparable with the calibration from the Tsai method.

One important thing to note in this scenario is that the observed pedestrian was not specifically tasked with traversing the scene for the sake of the calibration, and as such, they traverse only that portion of the scene that is the flat planar road surface. While it is important for the transformation of the camera system to the desired scene coordinate system that the person walks on a planar surface, it is not a requirement of the remainder of the calibration process that they do so. Rather, for the calibration process, it is important that point correspondences exist between views for as much of the extents of each view as is possible. The fourth camera covers the widest area of each of the cameras, and as such, the path traversed by the pedestrian does not cover a large extent of the view from camera 4. The result of this is that the calibration for camera 4 is poor relative to the other cameras, however, the same problem is observed for the Tsai calibration. It is also observable that the calibration starts to lose correctness once observations are made beyond the paved area. If a person were specifically tasked with traversing this scene, they could traverse the grassy areas which are non planar, which would improve the overall calibration (so long as the transformation to desired scene coordinate system only used those points where the person was on the planar surface).

camera	1	2	3	4
1	0.000	5.286	**3.110**	**6.418**
2	**7.022**	0.000	**6.375**	**10.139**
3	5.657	7.361	0.000	**6.088**
4	**23.719**	**15.659**	**8.935**	0.000

Table 3. Mean error, in pixels, between each camera, for the pedestrian calibration of the PETS2009 data set. Bold values indicate where this calibration produces a smaller error than the Tsai based calibration in Table 4.

camera	1	2	3	4
1	0.000	**5.214**	3.283	8.831
2	7.865	0.000	8.283	12.531
3	**5.182**	**7.237**	0.000	6.217
4	26.243	18.630	13.026	0.000

Table 4. Mean error, in pixels, between each camera, for the Tsai calibration of the PETS2009 data set. Bold values indicate where this calibration produces a smaller error than the pedestrian based calibration in Table 3.

camera	1	2	3	4
1	0.000	**2.714**	**11.117**	7.654
2	**11.280**	0.000	**17.994**	**9.775**
3	**2.682**	**1.675**	0.000	**3.080**
4	**5.247**	**2.939**	**8.092**	0.000

Table 5. Mean error, in pixels, between each camera, for the PETS 2006 pedestrian calibration. Bold values indicate where this calibration produces a smaller error than the Tsai based calibration in Table 6.

6.3. Further PETS calibrations

In the PETS 2006 and 2007 data sets it is more difficult to pick one single person that traverses a significant portion of the scene. However, the main part of the calibration process does not need to know that the person is a constant height, only that there are correlated points between camera views. The height of the person is used only for setting the scale, and as such there is no constraint that the same person must be used for all the head and foot obersvations so long as suitable care is taken for setting the scale. The relative errors in the calibration for each camera against each other camera, using both the pedestrian based calibration detailed in this paper as well as the Tsai calibrations provided with the data sets are shown in Tables 5 & 6 and 7 & 8. In each case, the pedestrian based calibration is again comparable in quality to the Tsai based calibration, and in some cases noticeably better.

camera	1	2	3	4
1	0.000	5.748	22.659	**7.186**
2	15.619	0.000	67.874	20.161
3	8.732	5.595	0.000	6.847
4	7.617	4.303	26.409	0.000

Table 6. Mean error, in pixels, between each camera, for the PETS 2006 Tsai calibration. Bold values indicate where this calibration produces a smaller error than the pedestrian based calibration in Table 5.

camera	1	2	3	4
1	0.000	**3.519**	**6.013**	6.605
2	**3.001**	0.000	**5.395**	6.955
3	**3.389**	**3.702**	0.000	**4.770**
4	8.289	8.194	12.234	0.000

Table 7. Mean error, in pixels, between each camera, for the PETS 2007 pedestrian calibration. Bold values indicate where this calibration produces a smaller error than the Tsai based calibration in Table 8.

camera	1	2	3	4
1	0.000	5.564	10.412	**6.379**
2	4.536	0.000	10.943	**5.975**
3	5.716	7.127	0.000	5.756
4	**5.275**	**7.113**	**11.603**	0.000

Table 8. Mean error, in pixels, between each camera, for the PETS2007 Tsai calibration. Bold values indicate where this calibration produces a smaller error than the pedestrian based calibration in Table 7.

7. Conclusion

The results presented here clearly suggest that using a pedestrian observation based approach to calibrating a network of surveillance cameras such as the one described in this paper can produce a useful calibration that is at least as good if not better than using a Tsai based approach that requires making measurements of the scene.

With respect to the calibration process, it should be observed that estimating the intrinsic parameters of the camera from scene data prior to bundle adjustment is extremely frustrating, especially with regards to the vansihing point approach, but experience shows that guessing the intrinsic parameters is sufficient to initialise the bundle adjustment and reach a useful calibration, so long as care is taken to slowly increase the complexity of the bundle adjustment, and not to throw all the point and camera data at the system in the first instance.

8. Acknowledgements

This work was partially funded by the EU FP7 project CO-FRIEND with grant no. 214975.[1]

[1]However, this paper does not necessarily represent the opinion of the European Community, and the European Community is not responsible for any use which may be made of its contents.

References

[1] S. Bougnoux. From projective to euclidean space under any practical situation, a criticism of self-calibration. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 790, 1998. 3

[2] J.-Y. Bouguet. www.vision.caltech.edu/bouguetj/calib_doc/. 1

[3] T. Chen, A. Del Bimbo, F. Pernici, and G. Serra. Accurate self-calibration of two cameras by observations of a moving person on a ground plane. In *Advanced Video and Signal Based Surveillance*, 2007. 1, 2

[4] J. M. Ferryman, editor. *Proceedings of the Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. IEEE, 2006. 5

[5] J. M. Ferryman, editor. *Proceedings of the Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. IEEE, 2007. 5

[6] J. M. Ferryman, editor. *Proceedings of the Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. IEEE, 2009. 2, 5

[7] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[8] R. Hartley and A. Zisserman. *Multiple view geometry in Computer Vision, 2nd Edtion*. Cambridge University Press, 2004. 1, 2, 3

[9] B. W. He and Y. F. Li. A novel method for camera calibration using vanishing points. In *Proceedings of the Mechatronics and Machine Vision in Practice conference*, 2007. 2

[10] E. E. Hemayed. A survey of camera self-calibration. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2003. 1

[11] N. Krahnstoever and R. S. Mendoça. Autocalibration from tracks of walking people. In *British Machine Vision Conference*, 2006. 1, 2, 3

[12] M. A. Lourakis and A. Argyros. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Trans. Math. Software*, 36(1):1–30, 2009. 4

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 20:91–110, 2003. 1

[14] F. Lv, Z. Tao, and R. Nevita. Self-calibration of a camera from video of a walking human. In *International Conference on Pattern Recognition*, 2002. 1, 2

[15] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment a modern synthesis. In *ICCV '99: Proceedings of the International Workshop on Vision Algorithms*, pages 298–372, 1999. 4

[16] R. Y. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 346–374, 1986. 1

[17] C. Zach. Simple sparse bundle adjustment. http://www.cs.unc.edu/ cmzach/opensource.html. 4