PTZ Camera Network Calibration from Moving People in Sports Broadcasts

Jens Puwein
ETH Zurich
puweinj@student.ethz.ch

Remo Ziegler
LiberoVision
http://www.liberovision.com

Luca Ballan, Marc Pollefeys
ETH Zurich
{luca.ballan,marc.pollefeys}@inf.ethz.ch

Abstract

In sports broadcasts, networks consisting of pan-tilt-zoom (PTZ) cameras usually exhibit very wide baselines, making standard matching techniques for camera calibration very hard to apply. If, additionally, there is a lack of texture, finding corresponding image regions becomes almost impossible. However, such networks are often set up to observe dynamic scenes on a ground plane. Corresponding image trajectories produced by moving objects need to fulfill specific geometric constraints, which can be leveraged for camera calibration.

We present a method which combines image trajectory matching with the self-calibration of rotating and zooming cameras, effectively reducing the remaining degrees of freedom in the matching stage to a 2D similarity transformation. Additionally, lines on the ground plane are used to improve the calibration. In the end, all extrinsic and intrinsic camera parameters are refined in a final bundle adjustment. The proposed algorithm was evaluated both qualitatively and quantitatively on four different soccer sequences.

1. Introduction

Pan-tilt-zoom (PTZ) camera networks are widely used in sports broadcasts. In order to analyze and understand the captured events, free-viewpoint video and augmented reality have proven to be valuable tools. The calibration of PTZ camera networks plays a crucial role in these applications, and it is a requirement for creating novel synthetic views of the recorded scenes. Even a small error in the alignment of two images can result in visible artifacts during scene re-rendering from a different point of view. This requires a joint calibration of the cameras and priority should be given to the minimization of this alignment error.

Typically, the calibration can only be performed based on the acquired footage since the access to the recording facilities is constrained by restrictive rules. Estimating the intrinsic and the extrinsic parameters of each camera at each time instant from the recorded videos is a challenging problem. The wide baselines and the absence of textured regions, typical for the footage recorded during soccer or rugby games, e.g., make standard calibration techniques hard to apply. While appearance based matching techniques succeed in relating contiguous images of a video captured by the same camera, in general, no correspondences can be found between images captured by different cameras. On the other hand, methods solely relying on the detection of field model lines work only if a model is available and a minimum number of lines is visible and detectable in each image.

While appearance based matching techniques are not applicable to find correspondences, player motions provide very strong cues. Just like feature correspondences, corresponding player tracks from different cameras need to fulfill specific geometric constraints.

In this paper, we present an approach to multi-view PTZ camera calibration of sports broadcasts. Since the camera network is set up to capture players moving on a ground plane, the player trajectories are leveraged for the camera calibration. If present, field lines are used to further improve the calibration. The resulting calibration technique shows to be robust and allows the calibration of challenging footage like the one obtained from soccer matches.

2. Related work

In sports broadcasts, it is common to use a 3D line model of a standardized playing field for calibration. Relating image lines and model lines leads to a set of 2D-to-3D line correspondences. Correspondences are hypothesized and the resulting camera parameters are verified by comparing the projected model lines with the image lines [8]. Once the center of projection of the camera is known, pan, tilt and zoom can be determined more robustly by exhaustively

comparing image lines with a database of projected model lines for different values of pan, tilt and zoom [23]. However, such approaches depend on a minimum number of visible lines and fail if this requirement is not met. If team logos, advertisements or other distinctive features are present on the playing field, calibration is still possible [20]. In sports like soccer or rugby, however, no logos or advertisements are available on the playing field.

For static camera networks, object trajectories have already been exploited for calibration. For instance, Stein treated the case of unsynchronized cameras with known intrinsic parameters [22]. The extrinsic parameters are found by estimating the temporal offset and the homography relating the trajectories of objects moving on a common ground plane. The approach proposed by Jaynes additionally deals with an unknown focal length per camera, but it requires the user to specify two bundles of coplanar parallel lines for each camera [13]. Assuming a planar scene and synchronized cameras, image trajectories are then warped onto 3D planes and the extrinsics are found by estimating a 3D similarity transformation between these planes. The work of Meingast *et al*. does not require objects to move on a ground plane [19]. Track correspondences are hypothesized and verified through geometric constraints to determine the correct essential matrices and temporal offsets between views. In the case of single static cameras, objects with known properties provide constraints for camera calibration. A common approach is to treat people as vertical poles of constant height [18, 14, 6]. A solution based on objects moving on non-parallel lines at constant speeds was presented by Bose and Grimson [3].

However, all these approaches do not address the problem of cameras which pan, tilt and zoom during the recording. Moreover, in sports broadcasts, assumptions like the vertical pole and the constant speed of the players cannot be used.

3. Algorithm Overview

The input to our problem consists of a number of synchronized video sequences captured by PTZ cameras. Let I_i^c denote the image captured by camera c at time i. The desired output is the full calibration of the PTZ camera network, i.e. the intrinsic parameters K_i^c, the distortion parameters k_i^c, the rotation matrices R_i^c, and the centers of projection C^c for all the cameras at each time instant. Let P_i^c denote the projection matrix of camera c at frame i, defined as $\mathrm{P}_i^c = \mathrm{K}_i^c[\mathrm{R}_i^c| - \mathrm{R}_i^c\mathrm{C}^c]$. The center of projection (COP) of each camera is assumed to be constant over time, but unknown. In addition, the pan axis is assumed to be always perpendicular to the playing field. To simplify the equations in this paper, the world coordinate system is chosen such that the xy-plane coincides with the ground plane. Hence, the pan axis is parallel to the z-axis, and the tilt axis is parallel to the xy-plane and orthogonal to the pan axis (see Figure 1). The pan angles, tilt angles and focal lengths of each camera change with every frame.

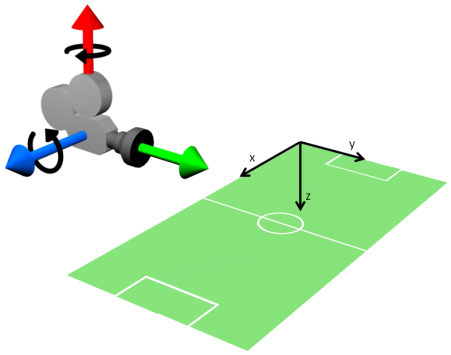


Figure 1. Typical camera arrangement with respect to the playing field. The pan axis (red) is orthogonal to the ground plane, while the tilt axis (blue) is parallel to it.

Ideally, the intrinsic parameters of the cameras would be pre-calibrated in a controlled environment, but access to the recording facilities and the cameras is typically restricted. We assume that the principal point lies in the center of the image. Since there is a correlation between the principal point and the remaining camera parameters, the error that is made by fixing the principal point in the middle of the image can be partially compensated by slightly changing the focal length, the COP and the orientation. Skew is assumed to be zero and the aspect ratios are assumed to be known. Under these assumptions, each intrinsic matrix K_i^c can be written as $diag(f_i^c, f_i^c, 1)$, where f_i^c denotes the focal length. Radial distortion is modeled with a single parameter k_i^c varying over time, such that

$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} = (1 + k_i^c(x_n^2 + y_n^2)) \begin{bmatrix} x_n \\ y_n \end{bmatrix},$$

where $(x_n, y_n)^T$ are the normalized image coordinates, and $(x_d, y_d)^T$ are the distorted ones.

First, each camera is calibrated independently. A local coordinate system is chosen for each camera, and the pan and tilt angles, the focal lengths and the distortion coefficients are determined with respect to this local coordinate system. Once these parameters are estimated, a metric rectification of the ground plane can be computed for each frame. A top-down view resulting from such a rectification is shown in Figure 2.

Subsequently, players are tracked in each video sequence independently and the tracks are warped onto the ground plane. The COPs and the rotation angles with respect to a common coordinate system are determined by the 2D similarity transformations relating the warped player tracks in all the video sequences. Camera parameters are refined in a final bundle adjustment.

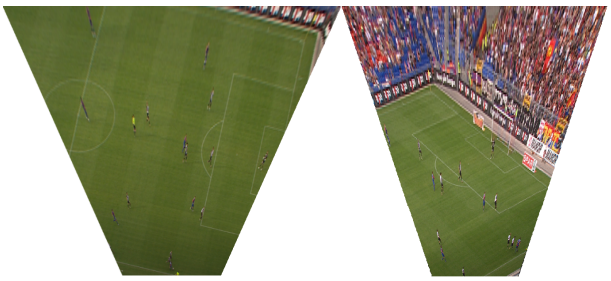


Figure 2. Top-down views of the playing field for two cameras.

4. Single Camera Calibration

In the first phase, each video sequence is treated independently. To improve readability, camera indices are removed in this section. SIFT features are extracted every ten frames and matched to SIFT features extracted ten frames before and after [16]. Additionally, Harris corners are extracted and tracked using a GPU implementation of the KLT tracker [17, 21]. Corresponding points in images I_i and I_j are related by the following homography [12]:

$$\mathrm{H}_{ij} = \mathrm{K}_j \tilde{\mathrm{R}}_j \tilde{\mathrm{R}}_i^T {\mathrm{K}_i}^{-1} = \mathrm{K}_j \mathrm{R}_{ij} {\mathrm{K}_i}^{-1}$$

where $\tilde{\mathrm{R}}_i$ and $\tilde{\mathrm{R}}_j$ are the rotation matrices defined in the local coordinate system of the camera (denoted by the tilde), at time i and j, respectively. R_{ij} is the relative rotation between the two views, which is independent of the coordinate system. A point x_j in image I_j is related to its corresponding point x_i in I_i as

$$\mathrm{x}_j \propto \mathrm{H}_{ij} \mathrm{x}_i$$

where x_i and x_j are given in homogenous coordinates and $\propto$ denotes equality up to scale.

From an image sequence produced by a rotating and zooming camera, it is in general possible to determine the intrinsic parameters of the camera as well as its relative rotations [11]. If only the focal lengths and the rotation angles are needed, there exists a minimal solution using 3 point correspondences between two frames [4]. We use this minimal solution to remove outliers in a RANSAC step and to get initial values for the focal lengths and the rotation angles [9]. The estimation of the focal length is unstable for small rotations, but the work of Agapito *et al*. suggests that the relative change of focal lengths can still be recovered [1]. This means that an absolute value has to be determined once from a pair of images separated by a sufficiently large rotation. Keeping the focal length fixed for those images, the remaining absolute values are given through the relative changes between frames. Given the initial absolute values for the focal lenghts, the initial rotations are determined.

The initial rotation angles and focal lengths are refined in a bundle adjustment using the sparseLM library together with the Cholmod solver (the same libraries are also used for all subsequent nonlinear least squares problems) [24, 15, 7]. The reprojection errors are minimized by optimizing for the feature positions, the rotation angles and the focal lengths. Under the assumptions stated in Section 3, depicted in Figure 1, the rotation matrix $\tilde{\mathrm{R}}_i$ is given by two rotation angles and can be rewritten as

$$\tilde{\mathrm{R}}_i = \mathrm{R}_{\alpha_i} \tilde{\mathrm{R}}_{\tilde{\gamma}_i}$$

where

$$\mathrm{R}_{\alpha_i} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_i & -\sin\alpha_i \\ 0 & \sin\alpha_i & \cos\alpha_i \end{bmatrix}$$

and

$$\tilde{\mathrm{R}}_{\tilde{\gamma}_i} = \begin{bmatrix} \cos\tilde{\gamma}_i & -\sin\tilde{\gamma}_i & 0 \\ \sin\tilde{\gamma}_i & \cos\tilde{\gamma}_i & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

While the tilt angle α_i is given with respect to a coordinate system common to all the cameras, the pan angle $\tilde{\gamma}_i$ is given in the local coordinate system of the respective camera, i.e., up to an additive constant c_γ. More precisely, $\tilde{\gamma}_i = \gamma_i + c_\gamma$. The relative rotation between frame i and frame j can be written as

$$\mathrm{R}_{ij} = \mathrm{R}_{\alpha_j} \tilde{\mathrm{R}}_{\tilde{\gamma}_j} \tilde{\mathrm{R}}_{\tilde{\gamma}_i}^T {\mathrm{R}_{\alpha_i}}^T = \mathrm{R}_{\alpha_j} \mathrm{R}_{\gamma_{ij}} {\mathrm{R}_{\alpha_i}}^T$$

which corresponds to

$$\begin{bmatrix} \cos\gamma_{ij} & -\sin\gamma_{ij}\cos\alpha_i & -\sin\gamma_{ij}\sin\alpha_i \\ \cos\alpha_j\sin\gamma_{ij} & \cdots & \cdots \\ \sin\alpha_j\sin\gamma_{ij} & \cdots & \cdots \end{bmatrix},$$

where γ_{ij} is the pan angle between frames i and j. α_j can be obtained as $\mathrm{atan2}(\mathrm{R}_{ij}(3,1), \mathrm{R}_{ij}(2,1))$. Similarly, α_i is obtained from $\mathrm{R}_{ij}(1,2)$ and $\mathrm{R}_{ij}(1,3)$. Care has to be taken when γ_{ij} is zero, i.e., when there is no panning motion. In this case, the tilt angles cannot be recovered. However, this is not an issue since normally for every frame in the sequence there exists at least another one with a different pan angle. Using the initial values of the absolute tilt angles α_i and the relative pan angles γ_{ij}, another bundle adjustment is performed, also including radial distortion coefficients k_i. Rotation matrices are now parametrized with pan and tilt angles instead of three Euler angles.

To further improve the results, image lines are included in the calibration process. After undistorting the input images, lines are extracted and matched between neighboring frames. For all sets of matching image lines, we try to find the camera parameters and the line positions on the ground plane which best explain the image lines.

In order to add lines to the calibration process, we introduce a mapping between points on the ground plane and image points. Let some frame r be a reference frame. At this

point, the projection matrix P_r is known up to the pan angle and the center of projection. The constraint of the ground plane being the xy-plane of the world coordinate system is not violated by arbitrary rotations around the z-axis, translations in x- and y-direction or scalings of the scene. Therefore, the world coordinate system can be chosen such that $\mathrm{P}_r = \mathrm{K}_r[\mathrm{R}_{\alpha_r}| - \mathrm{R}_{\alpha_r}(0,0,-1)^T]$. Since points on the ground plane are of the form $(x, y, 0)$, the homography mapping the ground plane to image I_i is $\mathrm{H}_{ri}\mathrm{K}_r\mathrm{R}_{\alpha_r}$.

The coordinates of a line on the ground plane are given by a point on the 3D unit sphere, parameterized by two rotation angles ρ and ϕ. Each image line is represented by point samples, where each point sample x corresponds to a point $\hat{\mathrm{x}}$ on a line $\mathrm{l}(\phi, \rho)$ on the ground plane. To specify the coordinates of $\hat{\mathrm{x}}$ lying on the line given by ρ and ϕ, we introduce an additional parameter λ for each point sample, such that $\hat{\mathrm{x}}$ is a function p of ρ, ϕ and λ:

$$\mathrm{x} \propto \mathrm{H}_{ri}\mathrm{K}_r\mathrm{R}_{\alpha_r}\hat{\mathrm{x}} = \mathrm{H}_{ri}\mathrm{K}_r\mathrm{R}_{\alpha_r}p(\rho, \phi, \lambda)$$

To extract lines, a Canny edge detector is used in the undistorted images (see Figure 3) [5]. E.g., white field lines on green background generate two lines. Such lines are very close and are merged into a single line. Correspondences between image lines are then determined by using the current estimates of the H_{ij}'s. If a line in image I_i is sufficiently close to a line in I_j after being transformed using H_{ij}, the two lines are considered a match. If, however, a third line is close to either one of those two lines, the match is simply discarded. For each set of matching lines, a line on the ground plane is initialized by calculating a least squares fit to the line samples projected onto the ground plane.

In addition to the reprojection errors obtained from the KLT tracks and the SIFT feature points, the reprojection errors of the line samples are added to the bundle adjustment. Additionally, e.g., on soccer fields, lines are either parallel or orthogonal. The deviation from known angles can be added as well. In a third bundle adjustment, we optimize for all pan angles, tilt angles, focal lengths, distortion coefficients and all the ϕ's, ρ's and λ's.

5. Camera Network Calibration

5.1. Fixing a World Coordinate System

So far, cameras were treated completely independently, and the projection matrices P_i^c were computed up to an unknown COP and up to an unknown offset of the pan angles.

The world coordinate system is chosen with respect to an arbitrary reference camera a and an arbitrary reference frame r, such that $\mathrm{C}^a = (0,0,-1)^T$ and $\gamma_r^a = 0$, i.e., $\mathrm{P}_r^a = \mathrm{K}_r^a[\mathrm{R}_{\alpha_r}^a| - \mathrm{R}_{\alpha_r}^a(0,0,-1)^T]$. For any additional camera b, we get $\mathrm{P}_r^b = \mathrm{K}_r^b[\mathrm{R}_r^b| - \mathrm{R}_r^b\mathrm{C}^b]$. The relative pan and translation with respect to camera a needs to be determined

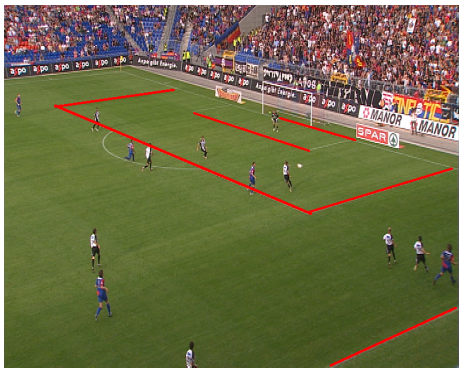


Figure 3. Lines detected in the undistorted image.

in order to find γ_r^b and C^b. By applying the transformation T_r^a with

$$\mathrm{T}_r^a = \begin{bmatrix} \mathrm{R}_{\alpha_r}^{a\ T} & (0,0,-1)^T \\ 0 & 1 \end{bmatrix}$$

to P_r^a and P_r^b we get $\mathrm{P}_r^a = \mathrm{K}_r^a\,[\mathrm{I}|0]$ and $\mathrm{P}_r^b = \mathrm{K}_r^b\,[\mathrm{R}|\mathrm{t}]$, where $\mathrm{R} = \mathrm{R}_r^b\mathrm{R}_{\alpha_r}^{a\ T}$ and $t = \mathrm{R}_r^b((0,0,-1)^T - \mathrm{C}^b)$. Since points in image I_i^c can be warped into a reference image I_r^c by using H_{ir}^c, the following formulas are derived for reference frame r and the index denoting the frame number is dropped to improve readability. P_c, K_c, R_c, R_{α_c} and R_{γ_c} denote P_r^c, K_r^c, R_r^c, $\mathrm{R}_{\alpha_r}^c$ and $\mathrm{R}_{\gamma_r}^c$. For camera matrices $\mathrm{P}_a = \mathrm{K}_a\,[\mathrm{I}|0]$ and $\mathrm{P}_b = \mathrm{K}_b\,[\mathrm{R}|\mathrm{t}]$, the homography H induced by the plane with normal n located at a distance d from the origin is given as [12]:

$$\mathrm{H} = \mathrm{K}_b(\mathrm{R} - \frac{1}{d}\mathrm{t}\mathrm{n}^T)\mathrm{K}_a^{-1}.$$

n is the normal of the ground plane, i.e., the xy-plane. Thus the associated normal vector is $(0,0,-1)^T$. However, since T_r^a was applied to the projection matrices, this normal needs to be transformed accordingly and we get

$$\mathrm{n} = \mathrm{R}_{\alpha_a}\begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}.$$

Because K_a, K_b and the tilt angles are known, the image coordinates and H can be transformed accordingly, leaving four unknowns γ_b, $\tilde{t_x}$, $\tilde{t_y}$ and $\tilde{t_z}$:

$$\begin{aligned}\tilde{\mathrm{H}} &= \mathrm{R}_{\alpha_b}^T\mathrm{K}_b^{-1}\mathrm{H}\mathrm{K}_a\mathrm{R}_{\alpha_a} \\ &= \mathrm{R}_{\alpha_b}^T\mathrm{K}_b^{-1}\mathrm{K}_b(\mathrm{R}_{\alpha_b}\mathrm{R}_{\gamma_b}\mathrm{R}_{\alpha_a}^T - \frac{1}{d}\mathrm{t}\mathrm{n}^T)\mathrm{K}_a^{-1}\mathrm{K}_a\mathrm{R}_{\alpha_a} \\ &= \mathrm{R}_{\gamma_\mathrm{b}} - \mathrm{R}_{\alpha_b}^T\frac{1}{d}\mathrm{t}\begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}^T = \begin{bmatrix} \cos(\gamma_b) & -\sin(\gamma_b) & \tilde{t_x} \\ \sin(\gamma_b) & \cos(\gamma_b) & \tilde{t_y} \\ 0 & 0 & \tilde{t_z} \end{bmatrix}\end{aligned}$$

Matrix $\tilde{\mathrm{H}}$ is defined up to scale, therefore a linear solution can be obtained from 2 correspondences of points lying on the ground plane, i.e., 4 equations.

One might think that 3 equations are enough to recover one angle and a translation up to scale. However, $\frac{1}{d}\mathrm{t}$ is unaffected by scale changes of the scene, hence it is not up to scale. Determining $\tilde{\mathrm{H}}$ can also be seen as finding the similarity transformation - scale, 2D translation and one rotation angle - between two top down views.

If the points are not lying on the ground plane, 3 point correspondences are still enough to determine pan and translation [10]. A method to extract correspondences between points on the ground plane is presented in the following subsection.

5.2. Correspondences From Player Trajectories

To generate player trajectories, we used a simple blob tracker. Foot positions are extracted from the player segmentations. A Gaussian mixture color model is created for both foreground and background colors by having the user select a few pixels with a few strokes [2].

The segmentation results in a few blobs. Associations between the blobs of different frames are made based on spatial proximity in a very conservative way. Whenever the trajectories of two blobs get too close, the tracking of these blobs stops. We use the strategy of Lv *et al*. to extract the foot positions of the players by using the principal axes of the blobs [18]. An example of foot tracks is shown in Figure 4. If a good people tracker is available, image trajectories can also be created from the centers of bounding boxes. However, the trajectories do not lie on the ground plane anymore, which has to be taken into account [10].

For each camera, image trajectories of the feet are then warped onto the ground plane given in the local coordinate system of the respective camera, as explained in Section 4. We define the set of candidate matches for each pair of cameras as all the possible pairs of extracted trajectories with a temporal overlap of at least 10 frames. Since the sought after transformation, a similarity transformation, is determined by two point correspondences, one correct trajectory match is sufficient to find the remaining unknowns in the absence of errors. This holds even if the tracks are perfectly straight. The only condition which has to be fulfilled is that the trajectory itself is not a single point, i.e., it is not related to a static object.

In order to find the correct association between player tracks, a similarity transformation $\mathcal{S}$ is generated for every candidate match and verified for all remaining candidates. The transformation leading to the largest number of inliers is assumed to be correct and the inliers represent matching trajectories. For a trajectory σ let σ_i denote the position at frame i. Let $\mathcal{S}(\sigma_i)$ denote the result of applying transformation $\mathcal{S}$ to point σ_i. A candidate match (σ, τ) is considered an inlier of transformation $\mathcal{S}$, if $\|\mathcal{S}(\sigma_i) - \tau_i\|_2 < t_{\text{dist}}$ for all overlapping frames i.

However, depending on the input footage, the accuracy of the self-calibration varies and the metric rectification can be inaccurate. Therefore, in some cases, trajectories are not aligned accurately enough to correctly determine the inliers. In order to compensate for such inaccuracies, we determine a homography from two candidate matches. A homography can explain any projective transformation between planes. Instead of testing all possible combinations of two candidate matches, the similarity transformations obtained before are used to determine suitable combinations. Whenever a second match is compatible with the similarity transformation obtained from the first match, the homography is calculated and verified. A candidate match is considered compatible with a transformation, if the shape and size of the transformed track is similar to the one of its match. Let $v_{ij}(\sigma)$ denote the vector from the position at frame i to the position at frame j of trajectory σ. Formally, for a candidate match (σ, τ) and a transformation $\mathcal{S}$, this means that the following two conditions need to be fulfilled:

$$\angle(v_{be}(\mathcal{S}(\sigma)), v_{be}(\tau)) < t_{\text{angle}}$$
$$\max\left(\frac{\|v_{be}(\mathcal{S}(\sigma))\|_2}{\|v_{be}(\tau)\|_2}, \frac{\|v_{be}(\tau)\|_2}{\|v_{be}(\mathcal{S}(\sigma))\|_2}\right) < t_{\text{length}},$$

where b denotes the first frame and e the last frame which σ and τ have in common. A valid set of correspondences is a 1-to-1 matching of image trajectories. If one track is matched to more than one other track, only the match with the lowest mean distance between corresponding points is used. Given the matching image trajectories, the homography induced by the ground plane can be calculated for every pair of images I_i^a and I_j^b, as explained in the previous subsection. To improve the calibration, additional features can be matched using these homographies. In soccer, the lines are the obvious choice. Detected line segments are matched if they are close enough but not too close to other segments. As for the self-calibration, image line samples are allowed to slide along the line on the ground plane. This time, a line on the ground plane is the same for all cameras and is given in the common world coordinate frame.

In a final bundle adjustment, all correspondences available are used to determine the full calibration of the camera network. This includes the camera intrinsics and extrinsics, one distortion coefficient per camera and image, SIFT and KLT feature positions extracted in the self-calibration stage, world positions of feet and heads as well as world positions of lines on the ground plane. The errors to be minimized are the reprojection errors and the deviations of the angles between lines from the known angles.

6. Evaluation

The proposed algorithm was evaluated on four different soccer game sequences. Each sequence was captured with two synchronized SD cameras working at a resolution of

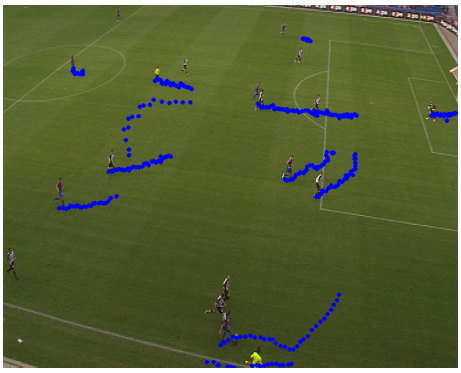


Figure 4. Foot tracks extracted from a soccer sequence.

720x576 pixels and undergoing a large range of rotations and zooms. The length of the sequences varies between 10 and 17 seconds, i.e., between 250 and 425 frames. The obtained results were evaluated both quantitatively and qualitatively.

Quantitative comparison In order to generate convincing synthetic views, a consistent calibration with respect to the different PTZ cameras is very important. To evaluate this property numerically, we used the root mean square of the symmetric epipolar transfer distances defined as the distance between a point in the first image and the epipolar line of its corresponding point in the second image and vice versa. More precisely, we evaluated

$$e_{\text{epipolar}} = \sqrt{\frac{1}{2N}\sum_{i=1}^{N}(d(\mathrm{x}'_i, \mathrm{F}\mathrm{x}_i)^2 + d(\mathrm{x}_i, \mathrm{F}^T\mathrm{x}'_i)^2)},$$

where F denotes the fundamental matrix, and $d(\mathrm{x}_i, \mathrm{F}\mathrm{x}_i)$ denotes the epipolar transfer distance defined as the Euclidean distance between point x_i and line $\mathrm{F}\mathrm{x}_i$. Salient corresponding points were selected manually over the course of the whole sequence. This includes heads, feet and joints of players, intersections of field lines, the corners of the goals and banners at the sidelines.

The symmetric epipolar transfer distance was evaluated on all the tested sequences. The calibration obtained by our method, once ignoring image lines and once accounting for image lines, was compared with a calibration obtained by manually selecting field model points. More precisely, this latter calibration was obtained by hand-clicking between 4 and 15 model points of the playing field every tenth frame. This includes all the corners and intersections formed by field lines, as well as the penalty points. Given the selected 3D-2D correspondences, the pan and tilt angles, the focal lengths and the COP were determined for each camera separately by minimizing the reprojection errors. Due to the limited number of hand-clicked points, radial distortion was not taken into account in the manual calibration.

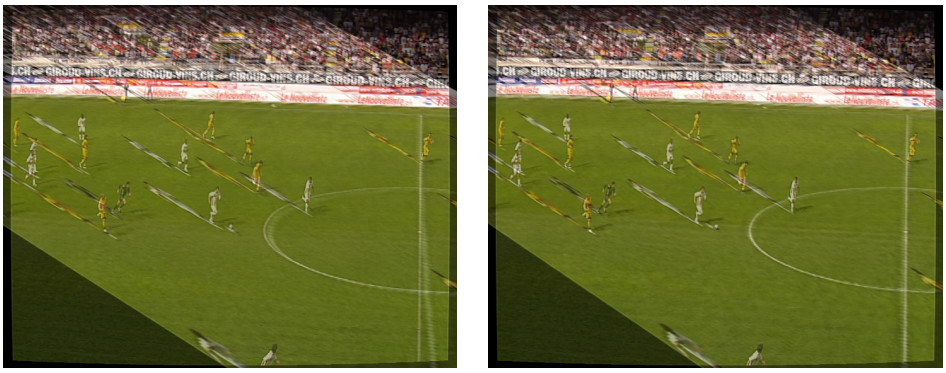
Table 1 summarizes the root mean square of the symmetric epipolar transfer distances obtained in the different sequences for both the manual calibration and our approach. Additionally, the table reports the median of all the computed epipolar transfer distances $d(\mathrm{x}_i, \mathrm{F}\mathrm{x}_i)$. The second column indicates the number of hand-clicked correspondences used to calculate e_{epipolar} for the manual calibration. Adding correspondences between field lines improved the results and in three out of four sequences, our approach leads to a smaller e_{epipolar} since it takes both cameras into account for the estimation. In sequence 4, the error obtained by our approach is a bit higher than the one obtained using manual calibration. This is due to the fact that in this sequence only few players are moving and the tracks of neighboring players got confused.

Figure 6 shows the distribution of the epipolar transfer distances $d(\mathrm{x}_i, \mathrm{F}\mathrm{x}_i)$ for all the points and sequences. It is visible that by using our method the distances concentrate more around zero. An additional comparison is shown in Figure 7. Here, tilt angles and focal lengths obtained after single camera calibration (Section 4), network calibration (Section 5), and by manual calibration for both cameras of sequence 1 are illustrated. The values for the manual calibration are not available every tenth frame because in some parts of the sequence not enough model points were visible. The shapes of the curves are very similar, suggesting that the estimations of the relative changes between frames are very similar. The additional constraints added by optimizing for two cameras simultaneously lead to a shift of the curves estimated from a single camera towards the ones estimated using manual calibration. Rotation angles and focal lengths obtained from single rotating cameras are correlated and, to some extent, increased focal lengths can be compensated by decreasing the rotation angles and vice versa [1].

Qualitative comparison Using the homography induced by the ground plane, the input image of one camera was warped and blended with the corresponding input image of the other camera. A result obtained accounting for lines and a result obtained without using any lines are shown in Figure 5. Non-overlapping lines and players cause visible artifacts when generating novel synthetic views using view-dependent textures. Hence it makes sense to minimize such errors.

7. Conclusion

In this paper, we presented an approach to multi-view PTZ camera calibration for sports broadcasts. Each camera is first calibrated in its own local coordinate system and then the player trajectories are leveraged to establish a common world coordinate system. Matching field lines further improves the calibration. Differently from previous approaches, the proposed calibration method does not rely on a 3D line model of the playing field or on detectable fea-



(a) (b)

Figure 5. Blend of two images aquired by two cameras observing the same scene. The first image was warped according to the homography induced by the ground plane and superimposed to the second one. Figure (a) shows the result obtained ignoring lines. Figure (b) shows the result obtained accounting for lines.

		manual calibration		our approach, no lines		our approach, incl. lines	
	#points	RMS error [px]	median [px]	RMS error [px]	median [px]	RMS error [px]	median [px]
sequence 1	108	2.9979	1.7343	2.8901	1.9629	1.6159	0.8144
sequence 2	147	6.2398	1.8590	2.6000	1.5034	1.8539	1.0943
sequence 3	55	2.8181	2.1985	4.8823	3.9737	1.3922	0.7909
sequence 4	82	2.5651	1.4776	8.5352	6.4020	4.5386	3.2121

Table 1. Epipolar transfer distance computed for the tested sequences.

tures on the field. This makes our method more robust in the absence of visible field lines and textured regions. The results presented in this papers show that the proposed technique can handle challenging sequences recorded by PTZ cameras separated by very wide baselines.

Limitations and Future Work Problems can be encountered in sequences where a camera constantly exhibits large focal lengths because the projection becomes almost orthographic and no perspective information can be inferred. In this case, self-calibration is difficult and typically inaccurate. However, as seen in Figure 7, if the camera zooms out at some point, the calibration can still be recovered.

In zoomed-in cameras, due to the narrow field-of-view, only a few players are visible, which means that there are only a few trajectories and hence trajectory matching becomes more difficult. Since a zoomed-in camera following the action needs to move fast, motion blur leads to inaccurate tracking results which, again, complicates trajectory matching. We would like to address these issues to be able to deal with constantly zoomed-in cameras and blurry subsequences.

If only a few of the trajectories extracted in the different cameras stem from common objects, the number of inliers obtained using the correct matching can be lower than the number of inliers obtained by a wrong matching. Additional cues like player colors can help in reducing the number of outliers.

8. Acknowledgments

The data is courtesy of Teleclub and LiberoVision. This project is supported by a grant of CTI Switzerland, the 4DVideo ERC Starting Grant Nr. 210806 and the SNF Recording Studio Grant.

References

[1] L. Agapito, E. Hayman, and I. Reid. Self-calibration of rotating and zooming cameras. *Int. J. Comput. Vision*, 2001.

[2] J. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, 1998.

[3] B. Bose and E. Grimson. Ground plane rectification by tracking moving objects. In *IEEE International Workshop on Visual Surveillance and PETS*, 2004.

[4] M. Brown, R. Hartley, and D. Nister. Minimal solutions for panoramic stitching. In *Proc. CVPR*, 2007.

[5] J. Canny. A computational approach to edge detection. *IEEE Trans. PAMI*, 1986.

[6] 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 *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS).*, 2007.

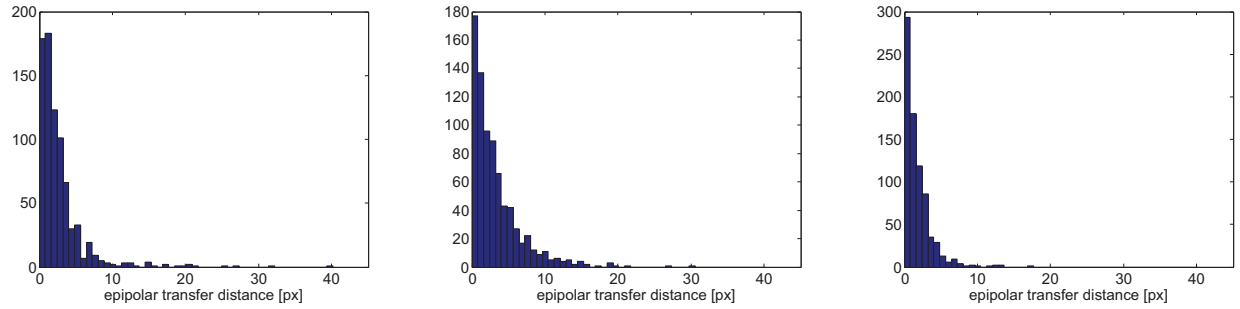


Figure 6. Distribution of the epipolar transfer distances for all sequences obtained by manual calibration (a), automatic calibration not accounting for lines (b) and automatic calibration including lines (c).

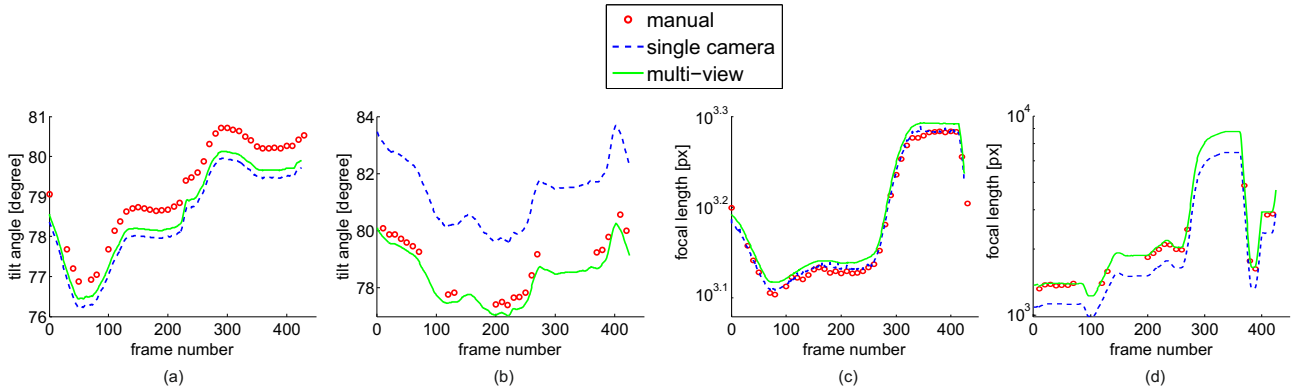


Figure 7. (a) and (b): tilt angles computed in each frame of sequence 1 for both cameras. (c) and (d): focal lengths computed in each frame of sequence 1 for the same cameras, shown at a logarithmic scale.

[7] Y. Chen, T. A. Davis, W. W. Hager, and S. Rajamanickam. Algorithm 887: Cholmod, supernodal sparse cholesky factorization and update/downdate. *ACM Trans. Math. Softw.*, 2008.

[8] D. Farin, S. Krabbe, W. Effelsberg, and P. H. N. de With. Robust camera calibration for sport videos using court models. In *SPIE Storage and Retrieval Methods and Applications for Multimedia*, 2004.

[9] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981.

[10] F. Fraundorfer, P. Tanskanen, and M. Pollefeys. A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In *Proc. ECCV*, 2010.

[11] R. I. Hartley. Self-calibration of stationary cameras. *Int. J. Comput. Vision*, 1997.

[12] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[13] C. Jaynes. Multi-view calibration from planar motion for video surveillance. In *Second IEEE Workshop on Visual Surveillance (VS'99)*, 1999.

[14] N. Krahnstoever and P. Mendonca. Bayesian autocalibration for surveillance. In *Proc. ICCV*, 2005.

[15] M. I. Lourakis. Sparse non-linear least squares optimization for geometric vision. In *Proc. ECCV*, 2010.

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 2004.

[17] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981.

[18] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *PAMI*, 2006.

[19] M. Meingast, S. Oh, and S. Sastry. Automatic camera network localization using object image tracks. In *Proc. ICCV*, 2007.

[20] J. Puwein, R. Ziegler, J. Vogel, and M. Pollefeys. Robust multi-view camera calibration for wide-baseline camera networks. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2011.

[21] J. Shi and C. Tomasi. Good features to track. In *Proc. CVPR*, 1994.

[22] G. P. Stein. Tracking from multiple view points: Self-calibration of space and time. *Proc. CVPR*, 1999.

[23] G. Thomas. Real-time camera tracking using sports pitch markings. *Journal of Real-Time Image Processing*, 2007.

[24] 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*, 2000.