CALIBRATING CAMERAS IN POOR-CONDITIONED PITCH-BASED SPORTS GAMES

Rui Zeng[1*] *Ruan Lakemond*[2] *Simon Denman*[1] *Sridha Sridharan*[1] *Clinton Fookes*[1] *Stuart Morgan*[3]

[1]Queensland University of Technology [2]Imagination Technologies [3]La Trobe University
{r5.zeng, s.denman, s.sridharan, c.fookes}@qut.edu.au
ruan.lakemond@imgtec.com s.morgan@latrobe.edu.au

ABSTRACT

Camera calibration is a preliminary step in sports analytics which enables us to transform player positions to standard playing area coordinates. While many camera calibration systems work well when the visual content contains sufficient clues, such as a key frame, calibrating without such information, such as may be needed when processing footage captured by a coach from the sidelines or stands, is challenging. In this paper an innovative automatic camera calibration system, which does not make use of any key frames, is presented for sports analytics. The proposed system consists of three components: a robust linear panorama module, a playing area estimation module, and a homography estimation module. It can eliminate distortion and calibrate the camera in each frame simultaneously, using correspondences between pairs of consecutive frames. Experiments on real data evaluate the performance and demonstrate the robustness of the system.

Index Terms— Homography estimation, sports video analytics, image stitching

1. INTRODUCTION

Calibrating cameras in sports scenes is often vital for further analysis such as player tracking, tactical and strategy analysis etc; as the camera calibration enables us to determine the real-world positions of players from their positions in a video.

In the past decade, the predominant approach has been the use of a key frame to calibrate cameras. Typically, a frame which has clear and complete line markings is set to be the key frame and then remaining frames are calibrated by using relationship between itself and the key frame. The early approach of [1] detected straight lines on the tennis court using color and local texture constrains and subsequently computed camera parameters by matching the intersections of detected lines with a standard court model. Hu *et al*. [2] improved on this method by adding color dominant segmentation to robustly determine the boundary of the basketball playing area. Lu *et al*. [3] detected the boundaries of the basketball court using a Canny detector [4] to calibrate a key frame and applied Iterated Closest Point (ICP) [5] to calibrate the remaining frames. Sha *et.al* [6] manually located reference points at the beginning of the swimming race frame. Then the calibration is performed by locating the reference points between adjacent frames using a SIFT feature extractor.

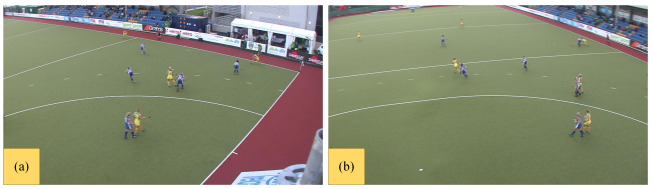

Fig. 1: Two challenging examples in camera calibration.

While many good methods have been proposed, a few commonly noticeable issues endure. They are related to the difficulty of finding, calibrating, and manually registering the key frame. A few examples are shown in Fig. 1. The first example only contains parts of the playing area due to the small field-of-view. The limited information makes it unsuitable as a key frame. While the second frame covers the whole playing area, the missing line on the farthest side of the playing area and the camera distortion limits it's use as a key frame. Such types of frames frequently appear in video clips which are captured by low-quality cameras positioned in the stands by a coach, or a spectator. Thus for such, there is no so-called key frame that can be used as a reference. These examples demonstrate a common problem: *when a video clip does not have a key frame, prior work generally fails due to the limited information provided by each individual frame*.

Aiming to address this problem, we present an automatic camera calibration system to calibrate each frame in a video clip which does not contain any key frames. The proposed system consists of three components: a robust linear panorama module, a playing area estimation module, and a homography estimation module. In addition, this system can undistort each frame while calibrating using correspondences between adjacent frames. We focus on field hockey in this work, but the technique is general and applicable to other team sports such as soccer, rugby, etc.

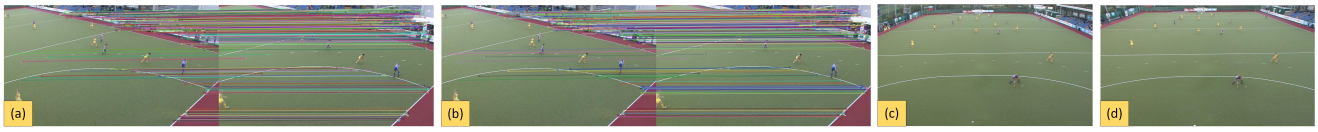

Fig. 2: (a) shows an initial correspondence matching completed by fast approximate nearest neighbor. (b) depicts the inliner correspondences refined by RANSAC. (c) and (d) are an original frame with lens distortion and its corresponding undistorted frame, respectively.

2. APPROACH

Our camera calibration system can be described as a combination of three interacting, but clearly separated modules: robust linear panorama generation (Sec. 2.1), playing area estimation (Sec. 2.2), and camera calibration (Sec. 2.3).

2.1. Robust linear panorama generation

We formulate the frame stitching and frame undistortion tasks as a combined panorama generation pipeline, enabling them to be completed simultaneously.

The first step is to extract feature points in each frame. SURF [7] is employed here to detect feature points. Subsequently, the DAISY [8] descriptor is used to describe dense information of the distribution around the detected feature points. The SURF detector and DAISY descriptor are chosen for two reasons: (1) the SURF key point detector is stable in local patches and invariant to scale and affine transforms. The perspective transform between every pair of adjacent frames is not significant and hence it can be regarded as an affine transform. (2) The DAISY descriptor is a dense feature extractor which can produce distinguishable features in a sparse area. Consider that what we are dealing with is a video sequence such that consecutive frames have a clear temporal order. Thus a probability model [9] is not employed here to estimate the correlation of one frame with every other frame. A frame is only registered with its adjacent frame in a linear way using a K-D tree.

Once the correspondences have been found, the correlation between them can be formulated by utilizing $\mathbf{c}'_{i,j+1} = \mathbf{H}'_{j,j+1}\mathbf{c}'_{i,j}$, where $\mathbf{H}'_{j,j+1}$ is the homography matrix; $\mathbf{c}'_{i,j}$ and $\mathbf{c}'_{i,j+1}$ are the homogeneous coordinates of the ith pair of correspondences that fall on two consecutive frames, j and $j+1$ respectively. Considering that matched correspondences may contain outliers, we make use of Random Sample Consensus (RANSAC) to refine $\mathbf{H}'_{j,j+1}$. Subsequently, a correspondence pair is labeled as an inlier or outlier according to the distance $\left\|\mathbf{H}'_{j,j+1}\mathbf{c}'_{ij} - \mathbf{c'}_{i,j+1}\right\| < \sigma$.

Since a single camera can suffer from severe camera distortion (particularly when wide-angle zoom lenses are used), the homography that we obtain from the above process may not be accurate, and a camera undistortion algorithm is required. A standard approach is to make use of camera intrinsic parameters and extrinsic parameters to analytically undistort frames. However, in sports videos captured from a single camera, camera parameters vary due to changes in zoom, and changes in the relative position between the playing area and the camera. To tackle this problem, a correspondences-based approach inspired by [10][11] is applied to undistort each frame. Instead of estimating camera intrinsic parameters and extrinsic parameters, we correct lens distortion in each frame using correspondences found between adjacent frames.

To simplify the lens distortion problem, we assume that the Center of Distortion (COD) is in the center of the frames. So the lens distortion model of a single frame j is defined as $\mathbf{d}_{ij} = \frac{1}{1+\lambda_j r_{ij}^2}\mathbf{d}'_{ij}$, where λ_j is the parameter of the lens division model, ${r_{ij}}^2$ is the squared distance between the feature point i and the COD, and $\mathbf{d}_{ij}$ is the undistorted coordinate of the ith pixel $\mathbf{d}'_{ij}$. Furthermore, the lens distortion parameter in two consecutive frames is assumed to be the same (while this is not strictly correct, the change between two consecutive frames is typically very small). Then the undistorted homography $\mathbf{H}_{j,j+1}$ which relates the two undistorted frames $\mathbf{c}_j$ and $\mathbf{c}_{j+1}$ can be defined as:

$$(\mathbf{D}_1 + \lambda_{j,j+1}\mathbf{D}_2 + \lambda_{j,j+1}\mathbf{D}_3)\mathbf{h}_{j,j+1} = 0, \tag{1}$$

where $\mathbf{c}_j$, $\mathbf{c}_{j+1}$ are the undistorted correspondences of $\mathbf{c}'_j$ $\mathbf{c}'_{j+1}$, $\mathbf{h}_{j,j+1}$ is the vectorized form of $\mathbf{H}_{j,j+1}$, $\lambda_{j,j+1}$ is the lens distortion parameter shared by the frame j and $j+1$, and $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{D}_3$ are the factor matrices. Let us assume that

$$\mathbf{A} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{I} \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} -\mathbf{D}_2 & -\mathbf{D}_3 \\ \mathbf{I} \end{bmatrix}, \ \mathbf{v} = \begin{bmatrix} \mathbf{h}_{j,j+1} \\ \lambda_{j,j+1}\mathbf{h}_{j,j+1} \end{bmatrix}, \tag{2}$$

such that the solution of $\lambda_{j,j+1}$ is found by iteratively solving $\mathbf{v}$ from

$$(\mathbf{A} - \lambda_{j,j+1}\mathbf{B})\mathbf{v} = 0, \tag{3}$$

and updating $\lambda_{j,j+1}$ by finding the smallest magnitude root of the scalar quadric equation [12][10]:

$$\mathbf{v}^\top(\mathbf{B}^\top + \lambda_{j,j+1}\mathbf{A}^\top)(\mathbf{A} - \lambda_{j,j+1}\mathbf{B})\mathbf{v} = 0. \tag{4}$$

In practice, a distortion parameter space $\{\lambda_{1,2}, \ldots, \lambda_{j,j+1}, \ldots, \lambda_{n-1,n}\}$ can be obtained from a video sequence containing n frames. The frame j makes use of $\lambda_{j,j+1}$ to do undistortion. The refined homography $\mathbf{H}_{j,j+1}$ between the frame j and $j+1$ is obtained from $\lambda_{j,j+1}$ using Eq. 1.

Once the refined homography between every two consecutive frames has been established, the homography between the anchor frame, n and the jth frame can be computed as

$H_{j,n} = H_{j,j+1}H_{j+1,j+2}\cdots H_{n-1,n}$. The coordinates in the anchor frame are selected as the coordinate system in the generated panorama. In practice, the anchor frame of the panorama is often selected as the last frame of a video clip.

To minimize the errors produced in linear matching, we employ the following error projection function to refine all of the homographies jointly:

$$e = \sum_j \sum_i \mathcal{L}_{ij} \left\| \mathbf{H}_{j+1,n}^{-1} \mathbf{H}_{j,n} \mathbf{c}_{ij} - \mathbf{c}_{i,j+1} \right\|_2^2 \quad . \quad (5)$$

Then the parameters in each homography are updated using the Levenberg-Marquardt (LM) algorithm [13].

Finally, the coordinate transformation between have generated panorama P and the frame I_j can be summarized as $\mathbf{c}_P = \mathbf{H}_{j,n}\mathbf{c}_{I_j}$. Fig. 2 shows results obtained in the process of robust linear correspondence matching. One may see that the outlier correspondences that usually appear on players have been filtered by RANSAC. Moreover, the camera distortion has been eliminated.

2.2. Playing area extraction

After generating a robust panorama from the video clip, the next step is to extract the playing area from this panorama. As outlined in [14], a color image I can be represented as a set of dominant colors and the percentage of occurrence of each. Thus color information can be used as an efficient low-level feature to segment the playing area from the background. Consider that the color of the playing area in pitch-like sports games is usually green and is typically much different to the surrounding area. The segmenation of the playing area can be done by searching for the green dominant color in the panorama. While there are a number of approaches for extracting dominant color [15][16][17], we use a dominant color segmentation approach using a combination of the CIELAB color space and Kmeans clustering. The reason why we chose CIELAB is that the chromaticity layer, 'A', indicates if the color falls along the red-green axis. As such, it may distinguish the green playing area from the surrounds. To improved the segmentation results, the small blobs and holes are filtered using morphological opening and closing operation respectively. Finally, the contour of the playing area is extracted by the Moore-Neighbor tracing algorithm and then smoothed using Savitzky-Golay filters [18].

2.3. Camera calibration

The task of camera calibration in the sports domain is to compute a geometric transformation $\mathbf{H}$ which maps a player location $\mathbf{p} = [x_\mathbf{p}, y_\mathbf{p}, 1]^\top$ in the image homogeneous coordinates to a point $\mathbf{p}' = [x_{\mathbf{p}'}, y_{\mathbf{p}'}, z_{\mathbf{p}'}, 1]^\top$ in the real-world homogeneous coordinates on the standard pitch plane P' [3][1], i.e. $\mathbf{p} = \mathbf{H}\mathbf{p}'$. Since the 3×3 matrix $\mathbf{H}$ has eight Degree of Freedom (DoF), four pairs of corresponding points are enough for estimating $\mathbf{H}$.

Taking the points on the detected contour of the playing area as input, a RANSAC-based line detector is used to parameterize all four boundary lines. Specifically, a line $\mathbf{l}$ is hypothesized by randomly selecting two points $\mathbf{m} = [x_\mathbf{m}, y_\mathbf{m}, 1]^\top$ and $\mathbf{n} = [x_\mathbf{n}, y_\mathbf{n}, 1]^\top$ which lie on the boundary lines of the segmented playing area. The idea of this method is to minimize the distance from the points in the boundary point set to the randomly hypothesized line $\mathbf{l}$; i.e.,

$$\min_{\substack{\mathbf{l} \\ [x,y,1]^\top \in \mathcal{P}}} \sum \mathbf{l} \times [x, y, 1]^\top. \quad \text{s.t.} \quad \mathbf{l} = \mathbf{m} \times \mathbf{n}, \quad \mathbf{m}, \mathbf{n} \in \mathcal{P}, \quad (6)$$

where $\mathcal{P}$ is the set of points $[x, y, 1]$ on the boundary lines, and $\times$ is the cross product of two vectors. After executing for a number of iterations, typically 10, the solution with the minimum distance is selected as a boundary line. To estimate other potential boundary lines, the points whose distance to the detected dominant line are less than T are eliminated and then we repeat the above method several times to get the remaining boundary lines.

Once the four corner points in the panorama $\{\mathbf{p}_i\}_{\mathrm{i}=1}^4$ have been determined by the intersections of the detected lines, the next step is to map them to the correspondences $\{\mathbf{p}'_i\}_{\mathrm{i}=1}^4$ that are in a standard field hockey model. Here, a Direct Linear Transformation (DLT) is applied to solve the homography from a set of equations:

$$\begin{bmatrix} \mathbf{0}^\top & -\mathbf{p'}_i^\top & y_{\mathbf{p}_i}\mathbf{p'}_i^\top \\ \mathbf{p'}_i^\top & \mathbf{0}^\top & -x_{\mathbf{p}_i}\mathbf{p'}_i^\top \end{bmatrix} \mathbf{h} = 0. \quad i = 1, \ldots 4 \quad , \quad (7)$$

where $\mathbf{h} \in \mathbb{R}^{9\times 1}$ is the vectorized homography $\mathbf{H}$. After mapping the panorama P to the standard pitch model P', the coordinates in frame I_j can be easily transformed to the standard pitch model using $\mathbf{c}_{P'} = \mathbf{H}\mathbf{H}_{j,n}\mathbf{c}_{I_j}$. Fig. 3 depicts the segmentation results and the panorama which is generated from a set of frames. It can be seen that there are many shadow players due to the image blending techniques. But the borders of the pitch are clear.

3. EVALUATION AND DISCUSSION

The proposed system was evaluated on four different field hockey video sequences. Each sequence was captured with a single camera at a resolution of 1920×1080, and includes a moderate range of rotations and zooms. The length of the video sequences vary between 10 and 18 seconds, i.e., between 250 and 450 frames. All of the video sequences sweep the whole playing area to ensure that the videos contain enough information for camera calibration.

Table 1 summarizes the average number of correspondences for each frame and the homography transfer error in terms of four video sequences used in our experiments. The average number of correspondences on each frame in a video indicates the performance of the feature extraction method in

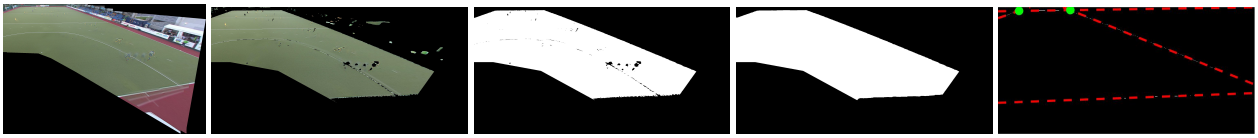

Fig. 3: leftmost shows the panoraa generated from a set of consecutive frames. The next three pictures depict the segmentation result, binarized playing area, and filtered playing area, respectively. The last image shows smoothed boundary lines and detected border lines.

Fig. 4: The calibrated results of representative frames in the panorama. Their corresponding regions are shown by yellow, magenta, cyan, red, green, blue, white and black box in the leftmost image. Meanwhile, the playing areas of these frames are drawn by the same color in a standard field hockey model in the second image.

Method	Metric	Video Sequences			
		1	2	3	4
DAISY+SURF	points	**264**	**288**	**281**	**272**
	RMS	**2.519**	**2.318**	**2.732**	**2.556**
KLT	points	34	28	29	30
	RMS	3.807	4.231	3.247	3.334
SIFT	points	92	74	81	87
	RMS	4.986	5.223	5.031	5.274

Table 1: The homography transfer error and the average number of the correspondences for the tested video sequences.

the panorama generation. The homography transfer error is computed by Root Mean Square (RMS) error which reports the pixel-level distance between the initially matched correspondences and the correspondences obtained from bundle adjustment. In addition, two commonly used feature extraction methods such as KLT [19] and SIFT are utilized to compare with our method. One may see that both the RMS and the average number of the points in our method outperforms that of the other two methods.

Fig. 4 shows some representative frames and the calibration result of these frames. The calibration of these frames can be derived from the panorama. One may see that the perspective distortion of the panorama is eliminated and the calibrated panorama is aligned to the standard field hockey model. Thus the positions on the panorama can be properly transformed to the standard field hockey model. We can see that the near end of the field cannot be captured by our model. This is a result of the zoom level of the camera leading to the it's effective position being in front of goal-line. As our technique is mapping images to a plane in front of the camera, we can only capture the portion of the field in a 180 degree arc in front of the camera.

We compare the proposed system to previous representative approaches like [1–3]. [1] focuses on tennis, where the smaller pitch size means that all line markings can be fully and clearly seen by a single camera. Each frame can be easily calibrated using the prior knowledge of the court lines. [2] calibrates the frame to the standard basketball court by detecting specific feature points on the court. These two methods rely on extraction of specific pitch structures on the playing area. However, each individual frame (Fig. 4: cyan, magenta, etc. frames) in our dataset does not contain enough clues for calibration in this manner due to the coverage of the camera. As such these systems [1, 2] are not applicable for our case. Regarding [3], it manually specifies the homogrpahy in the first frame of the video and derives other frames' location using ICP. In contrast, our system can automatically locate the frame relative to the pitch and undistort frames simultaneously. For some particular frames (Fig. 4: Black, Red, etc. frames), where the whole structure of the pitch is mostly visible, these methods [1–3] will still fail due to distortion and the disappearance of land marks in the far field of the camera. Examining Fig. 4, the resultant panorama does contain some small visual artifacts that are caused by small misalignments in the panorama construction. The misalignments propagate to the final perspective mapping. Compared to a semi automatic analytical such as [3], the proposed approach is less accurate, however it is fully automatic and places no requirements on the coverage of individual frames.

4. CONCLUSIONS

In this paper, an innovative camera calibration system for visual content without the need for a key frame is proposed. This system calibrates each individual frame through calibrating the panorama generated from these frames. It utilizes the holistic structure of the frames in a video rather than searching a key frame in isolation. The experiments showed that our system can work even if a given video does not contain any one frame that captures the entire playing area.

5. REFERENCES

[1] Jungong Han, Dirk Farin, and Peter HN de With, "Broadcast court-net sports video analysis using fast 3-d camera modeling," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1628–1638, 2008.

[2] Min-Chun Hu, Ming-Hsiu Chang, Ja-Ling Wu, and Lin Chi, "Robust camera calibration and player tracking in broadcast basketball video," *IEEE Trans. on Multimedia*, vol. 13, no. 2, pp. 266–279, 2011.

[3] Wei-Lwun Lu, Jo-Anne Ting, James J Little, and Kevin P Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Trans. on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1704–1716, 2013.

[4] John Canny, "A computational approach to edge detection," *IEEE Trans. on pattern analysis and machine intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[5] Zhengyou Zhang, "Iterative point matching for registration of free-form curves and surfaces," *International journal of computer vision*, vol. 13, no. 2, pp. 119–152, 1994.

[6] Long Sha, Patrick Lucey, Sridha Sridharan, Stuart Morgan, and Dave Pease, "Understanding and analyzing a large collection of archived swimming videos," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *European Conference on Computer Vision (ECCV)*, 2006.

[8] Engin Tola, Vincent Lepetit, and Pascal Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 815–830, 2010.

[9] Matthew Brown and David G Lowe, "Automatic panoramic image stitching using invariant features," *International journal of computer vision*, vol. 74, no. 1, pp. 59–73, 2007.

[10] Ruan Lakemond, Clinton Fookes, and Sridha Sridharan, "Practical improvements to simultaneous computation of multi-view geometry and radial lens distortion," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2011.

[11] R Steele and Christopher Jaynes, "Overconstrained linear estimation of radial distortion and multi-view geometry," in *European Conference on Computer Vision (ECCV)*, 2006.

[12] Gregory Boutry, Michael Elad, Gene H Golub, and Peyman Milanfar, "The generalized eigenvalue problem for nonsquare pencils using a minimal perturbation approach," *SIAM Journal on Matrix Analysis and Applications*, vol. 27, no. 2, pp. 582–601, 2005.

[13] Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[14] Junqing Chen, Thrasyvoulos N Pappas, Aleksandra Mojsilovic, and Bernice E Rogowitz, "Adaptive perceptual color-texture image segmentation," *IEEE Trans. on Image Processing*, vol. 14, no. 10, pp. 1524–1536, 2005.

[15] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin, "Color invariants for person reidentification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[16] Jesús Chamorro-Martínez, P Martínez-Jiménez, and José Manuel Soto-Hidalgo, "A fuzzy approach for retrieving images in databases using dominant color and texture descriptors," in *International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2010.

[17] Peng Wang, Dongqing Zhang, Jingdong Wang, Zhong Wu, Xian-Sheng Hua, and Shipeng Li, "Color filter for image search," in *ACM Multimedia (ACMMM)*, 2012.

[18] Michael Suhling, Muthuvel Arigovindan, Patrick Hunziker, and Michael Unser, "Multiresolution moment filters: Theory and applications," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 484–495, 2004.

[19] P. C. Wen, W. C. Cheng, Y. S. Wang, H. K. Chu, N. C. Tang, and H. Y. M. Liao, "Court reconstruction for camera calibration in broadcast basketball videos," *IEEE Trans. on Visualization and Computer Graphics*, vol. 22, no. 5, pp. 1517–1526, May 2016.