Accurate and Stable Camera Calibration of Broadcast Tennis Video

[1]Xinguo Yu, [2]Nianjuan Jiang, and [2]Loong-Fah Cheong

[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, xinguo@i2r.a-star.edu.sg
[2] Dept of Electrical and Computer Engineering, National University of Singapore, Singapore 117543, {u0308314, eleclf}@nus.edu.sg

Abstract: This paper presents an original algorithm for accurate and stable camera calibration of broadcast tennis video (BTV). That frame-data of BTV is often erroneous results in wildly fluctuating camera parameters. To meet this challenge, we propose a *frame grouping* technique, which groups frames together according to camera viewpoint. We then use a group-wise data analysis to obtain more stable parameters. Recognizing the fact that some of these parameters do vary somewhat even if they have a similar camera viewpoint, we further employ a *Hough-like search* to tune them, maximizing the reprojection similarity. This two-tiered process gains stability of the camera parameters, and yet ensures large reprojection similarity via the tuning step. The experimental results show that our algorithm is able to acquire accurate camera matrix.

Keywords: Sports Video, Camera Calibration, Group-Wise Data Analysis, Hough-like Search.

1. INTRODUCTION

In the field of sports video, the topic of camera calibration has attracted big attention because camera calibration can be used in a wide spectrum of applications such as arbitrary view presentation, 3D virtual content insertion [9], semantic analysis [7], and computer-aided refereeing [6].

A camera can be parameterized by both an intrinsic and an extrinsic model, which can be calibrated independently or simultaneously. The independent calibration generally has lighter computation burden, but it may easily cause error propagation from the intrinsic model to the extrinsic model. Accordingly, most researches focus on integrated methods that calibrate the intrinsic and the extrinsic model simultaneously. This is often accomplished by minimizing different kinds of cost functions, usually emphasizing on the reprojection similarity in the image space. Various solutions have been proposed, such as those based on gradient descent [4], interval analysis [2], etc. Unfortunately, these methods generally suffer from poor convergence, susceptibility to getting trapped in local extrema, or slow convergence. Camera calibration algorithms also differ in whether the model considers distortions or not, and in what ways the distortions are modeled [8]. Various camera calibration techniques are based on planar reference objects [10]. Feature points on a plane appearing in different multiple views are required for such plane-based calibration method. In broadcast tennis video (BTV), however, view changes occur within a very small range (less than 15^o most of the time).

There have been several papers which address camera calibration in the sports video domain. These algorithms used classical calibration techniques as opposed to self-calibration techniques [5]. The camera projection matrix is computed via solving a set of linear equations [1,6,7]. Such camera algorithms face various challenges such as inaccurate and incomplete features. In the case of BTV, camera calibration algorithms also have to deal with additional characteristics:

- Image features in tennis video span a small spatial volume because the two net poles are short.
- Camera recording BTV often pans and zooms, being mounted on a tripod. Thus algorithms assuming fixed intrinsic parameters cannot be used due to zooming changes the intrinsic parameters.

In this paper, we address these challenges in two phases. First, we use a Hough-like search to tune the initial features obtained by finding the intersection points of straight lines. We further exploit an extra constraint that is available in many sports videos, namely, a particular camera's view of the playfield reappears in many frames. We propose a grouping technique to leverage on this fact so as to combat against the effects of image noises. The by-product of this grouping is that frames with erroneous features are singled out because they probably do not belong to any group. The frames in the same group share similar values for some of the camera parameters, termed as the *group-invariant* parameters. The accuracy of these parameters can be improved by group-wise data analysis. Next, we use Hough-like search to refine some of these parameters, as in reality, they vary somewhat even though the camera is roughly having the same viewpoint. By adopting this two-phase process, our algorithm achieves both stability and accuracy of camera calibration.

The rest of this paper is organized as follows. Section 2 presents the proposed algorithm. Section 3 presents the experimental results. We conclude the paper in section 4.

2. CAMERA CALIBRATION ALGORITHM

This section presents the proposed two-phase algorithm using grouping and Hough-like search techniques. It aims to acquire accurate and stable camera matrix for each clip. Here, a clip refers to a sequence of consecutive frames shot by the same camera from a location around the tennis court.

2.1. Overview of Camera Calibration Algorithm

2.1.1. Projection Geometry

Fig 1 shows the Euclidean transformation between the real world and the image space for tennis video. The real-world point is represented by *w,* a homogenous 4-vector $(X, Y, Z, 1)^T$, *m* for the image point represented by a homogenous 3-vector $(x, y, 1)^T$, and P for the 3×4 camera projection matrix. Then for a pinhole camera, the mapping between the 3D world and the 2D image is written compactly as

$$m \cong \mathrm{P}w \qquad (1)$$

where $\cong$ means that two sides can differ by an arbitrary scale.

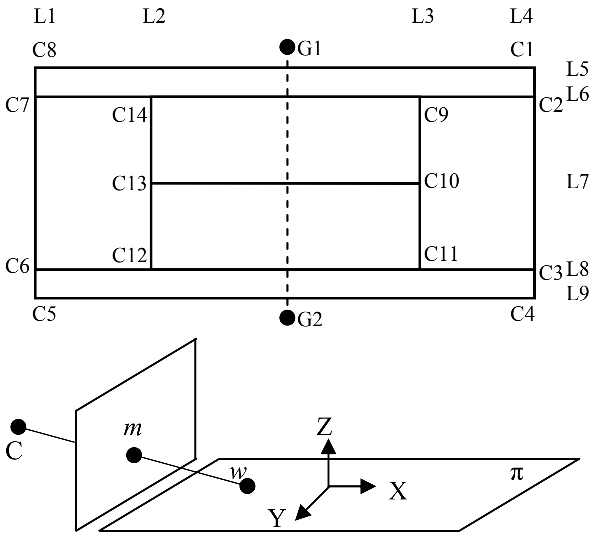

Fig 1. The Euclidean transformation between the real world and the image space for tennis video. The tennis court in the upper part of the figure appears on the plane π in the lower part of the figure.

2.1.2. Structure of Camera Calibration Algorithm

Among 11 camera parameters, some of them are group-invariant (GI), i.e. they are more or less constant in a group. Some are clip-invariant, abbreviated as CI; their values are unchanged over the entire video clip. More explanations of these terms are given in Section 2.3.3.

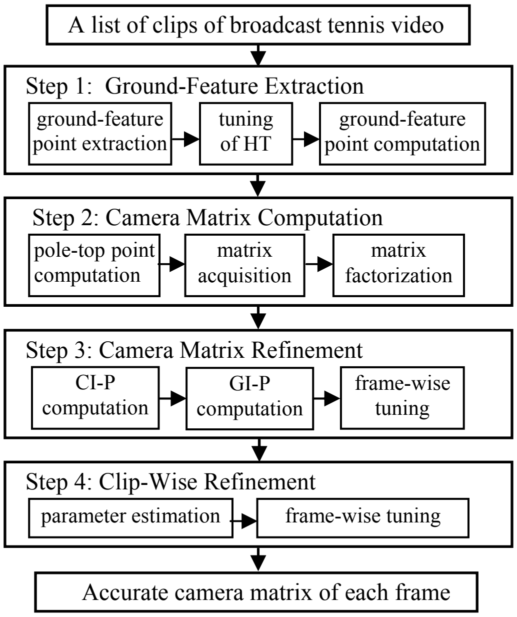
A list of clips of broadcast tennis video
Step 1: Ground-Feature Extraction
ground-feature point extraction
tuning of HT
ground-feature point computation
Step 2: Camera Matrix Computation
pole-top point computation
matrix acquisition
matrix factorization
Step 3: Camera Matrix Refinement
CI-P computation
GI-P computation
frame-wise tuning
Step 4: Clip-Wise Refinement
parameter estimation
frame-wise tuning
Accurate camera matrix of each frame

Fig 2. Block diagram of the proposed camera calibration algorithm. The input is a list of clips of broadcast tennis video and the output is the camera matrices of all the frames.

As depicted in Fig 2, the proposed algorithm takes four steps to obtain the accurate camera parameters. The first step finds the accurate ground points via a Hough-like search for the best matching homography transform (HT) starting from the intersections of straight lines. The second step obtains the initial camera matrix for each frame. This step first finds the two pole-tops via another Hough-like search. Then it computes the camera matrix and factorizes the matrix computed to obtain the 11 camera parameters for each frame. The third step refines the camera matrix computed. It first finds the clip-invariant parameters (CI-P) via finding the cluster centers of all their instances in a clip. Then it classifies all frames into groups according to their lookats and focal lengths. Then it obtains a better estimate of the group-invariant parameters (GI-P) via a group-wise data analysis. Except for the camera center, the rest of the group-invariant parameters are further refined via another Hough-like search, i.e. frame-wise tuning. The frame grouping in the third step also singles out frames that are not in any group. The last step estimates the camera projection matrices for those frames singled out in the third step, according to the values of their neighboring frames. Finally except for the CI parameters and the camera center, the rest of the estimated parameters are further tuned like those in step 3.

2.2. Ground-Feature Extraction

2.2.1. Initial Ground Feature Extraction

We first use a procedure to segment image court, in which only pixels on straight lines consisting of image court are kept and all other pixels are painted in court color. Then we use Hough transform to find all straight line equations and fit the lines detected into the court (*wire*) model. Once we have fitted the ground model of the tennis court we can find the image coordinates of the points C1 to C14 shown in Fig 1 by finding the intersection points of the straight lines detected. These form our initial ground features.

2.2.2. Ground Feature Refinement

Initial Homography: With the ground features extracted, we can determine the initial homography H in equation (2), which relates a point $w = (X, Y,\ 1)$ on the ground plane in the 3D world to an image point $m = (x, y, 1)$.

$$m \cong Hw. \qquad (2)$$

Similarity Measure of Two Court Images: Let C_{std} and C_{com} be any two images of the tennis court. We use L1 to L9 to denote the straight lines as shown in Fig 1. We use Li_{std} and Li_{com} to denote the sets of all the points contained in the straight line Li in C_{std} and C_{img}, respectively for i = 1 to 9. The measure function $M_L(i)$ of straight line Li is defined as

$$M_L(i) = \frac{|Li_{std} \cap Li_{com}|}{|Li_{std}|} \times 100. \qquad (3)$$

where $|\bullet|$ is the cardinality of a set**.**

The measure function $M_{cot}(C_{std}, C_{com})$ for measuring the similarity of the two courts is defined as.

$$M_{cot}(C_{std}, C_{com}) = \frac{1}{9}\sum_{i=1}^{9} \psi(i) . \qquad (4)$$

where $\psi(i) = M_L(i)$ if $M_L(i) > \beta(threshold)$; otherwise = 0.

Measure Function of Homography: Let C_{grd} be the ground model of the tennis court in the world coordinate system. We use $C_{hom}(H, C_{grd})$ to denote the transformed court from the ground model of the tennis court onto the image plane according to H. Let C_{img} be the segmented court in frame F.

With the two court images C_{img} and C_{hom}, we define the measure function $M_1(H, F)$ for a given H and a frame F as

$$M_1(H, F) = M_{cot}(C_{hom}, C_{img}). \qquad (5)$$

Homography Tuning: The following tuning procedure (Procedure 1) will significantly improve the match value. Let H_0 be the initial homography of the frame F. Then we prepare a small Hough space H_{sp} enclosing it. We compute $M_1(H, F)$ for all H in H_{sp} and adopt the best homography.

Procedure 1

```
Step 0: Input initial homography $H_0$ and segmented court.
Step 1: Form the Hough space $H_{sp}$ enclosing $H_0$;
Step 2: Initialize $V_{max} = M_1(H_0, F)$;  $H_{td} = H_0$;
        For all H in $H_{sp}$ do
            If $M_1(H, F) > V_{max}$, then $V_{max} = M_1(H, F)$; $H_{td} = H$ .
Terminate and output $H_{td}$.
```

Ground Feature Computation: Once we have obtained the best homography H_f for a frame F we re-compute all of the 14 ground features using H_f according to equation (2).

2.3. Camera Matrix Computation

2.3.1. Pole-Top Extraction

Here we aim to find the image coordinates of the two pole-tops to sub-pixel accuracy. Let (x_{g1}, y_{g1}) and (x_{g2}, y_{g2}) be the coordinates of G1 and G2 in a frame F, where G1 and G2 are the ground points of two poles. Let $(X_{g1}, Y_{g1}, 0)$ and $(X_{g2}, Y_{g2}, 0)$ be the coordinates of G1 and G2 in the 3D world. We define $w_{g1} = (X_{g1}, Y_{g1}, 1)$ and $w_{g2} = (X_{g2}, Y_{g2}, 1)$. Let H_F be the homography obtained in section 2.2. Then the formula $(I_{gi}, J_{gi}, K_{gi}) = H_F w_{gi}$, $x_{gi} = I_{gi}/K_{gi}$, $y_{gi} = J_{gi}/K_{gi.}$ (i= *1, 2*) is used to obtain their corresponding image positions.

Once we have computed (x_{g1}, y_{g1}) and (x_{g2}, y_{g2}), we can search for the two pole-tops L and R along the vertical poles in the image. Let $L_0 = (x_{g1}, y^0_{g1})$ and $R_0 = (x_{g2}, y^0_{g2})$ be the coordinates of the two extracted pole-tops in the image, being used as initial coordinates of the pole-tops in the following tuning procedure.

2.3.2. Pole-Top Tuning

Measure Function for Camera Matrix: Let C_{phy} be the complete model of the tennis court that includes both the ground model and the two net poles. We use $C_{pjd}(P, C_{phy})$ to denote the projected court from C_{phy} via P according to equation (1). P will be computed based on {C1, C2, …, C14} ∪ {L, R} (see next subsection). With the obtained P, we can formulate, for a given image F, the measure function $M_2(L, R, F)$ to tune the coordinates of the pole tops.

$$M_2(L, R, F)= M_{cot}(C_{pjd}, C_{img}). \qquad (6)$$

Based on the initial coordinates of the pole-tops, we create a Hough space H_{pole} = { (C1, C2, …, C14) ∪ $(x'_{G1}, y^0_{G1}+\sigma)$ ∪ $(x'_{G2}, y^0_{G2}+\tau)$, with σ and τ varying from -2 to 2 pixels in a step size of 0.01}. Then we use the following Procedure 2 to find the best coordinates of the two pole-tops in subpixel accuracy.

Procedure 2

```
Step 0: Input initial coordinates $L_0$ and $R_0$ and segmented court.
Step 1: Form the Hough space $H_{pole}$;
Step 2: Initialize $V_{max} = M_2(L_0, R_0, F)$;  $L_{td} = L_0$ and $L_{td} = L_0$;
        For all L and R in $H_{pole}$ do
            If $M_2(L, R, F) > V_{max}$, then $V_{max} = M_2(L, R, F)$;  $L_{td} = L$
            and $R_{td} = R$.
Terminate and output $L_{td}$ and $R_{td}$.
```

2.3.3. Camera Matrix Factorization

From $m \cong Pw$ (see section 2.1), we can solve for P using the direct linear transform method [4]. Next, to obtain the intrinsic and the extrinsic parameters, the projection matrix P can be factorized as

$$P = KR[I \mid -C]. \qquad (7)$$

where I, K and R are identity, calibration, and matrixes respectively, and C is the camera center. K is an upper triangular matrix encoding the intrinsic parameters:

$$K= \begin{bmatrix} f\gamma & s & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix}. \qquad (8)$$

where f is *focal length*, γ *aspect ratio*, s *skew* factor, and (u_0, v_0) coordinates of the *principal point*.

2.4. Camera Matrix Refinement

The frames sharing the same lookat and focal length are gathered in a group and they should have the same focal length f, principal point (u_0, v_0), camera center C, and the three angles for rotation (θ_x, θ_y and θ_z). These parameters form the group-invariant parameters. The remaining two parameters, namely the aspect ratio γ and the skew factor s are clip-invariant, particularly s=0. Note that we allowed the principal point (u_0, v_0) and the camera center C to vary with f, and they are thus not clip-invariant.

2.4.1. Frame Grouping

In broadcast tennis video (BTV), the lookat of a frame is usually some point on the court ground. Thus, its position in the world coordinate system can be computed by the homography obtained in Section 2.2.2. More exactly, the homography from the image space to the 3D world space is given by:

$$w = H_{trw}\, m \qquad (9)$$

where $H_{trw} = H^{-1}$. Let m_{ctr} = (w/2, h/2, 1) where (w/2, h/2) is the image center. Then $H_{trw}\, m_{ctr}$ is the lookat of the frame. The focal length of the frame is obtained by factorizing the camera projection matrix as shown in equation (8).

2.4.2. Clip- and Group-Invariant Parameter Computation

Clip-invariant parameters: Since the aspect ratio γ is constant in a clip, we find γ via gathering all their instances in a video clip and obtain their cluster centers.

Group-invariant parameters: To realize the Hough-like tuning of camera matrix, we define a measure function based on the 11 straight lines, that is, the nine ground lines and the two poles. The measure function $M_3(P, F)$ for the camera projection matrix P is defined as

$$M_3(P, F) = \frac{1}{11}\sum_{i=1}^{11}\psi(i) \qquad (10)$$

where $\psi(i)$ is the function defined in equation (4) and $\psi(10)$ and $\psi(11)$ are for the left and the right poles.

Tuning Procedure: For a frame F, let the camera projection matrix computed from its feature points be denoted by P_f. Then the initial camera projection matrix P_0 is produced by replacing the group-invariant parameters of P_f with the corresponding ones of the group's cluster center. We use *Procedure 3* to tune P_0**.** The Hough space H_{tune} centered at P_0 with the group-invariant parameters varying in a given step size around the center. We compute $M_3(P, F)$ for all P in H_{tune} and choose the best camera projection matrix.

Procedure 3

```
Step 0: Input initial camera matrix $P_0$ and segmented court.
Step 1: Form the Hough space $H_{tune}$ based on $P_0$.
Step 2: Initialize $V_{max} = M_3(P_0, F)$;  $P_{td} = P_0$.
        For all P in $H_{tune}$ do
            If $M_3(P, F) > V_{max}$, then $V_{max} = M_3(P, F)$; $P_{td} = P$.
        Terminate and output $P_{td}$.
```

3. EXPERIMENTAL RESULTS

We compare the developed algorithm in this paper (*shorted as ours*) with Tsai's algorithm presented in [8] (*shorted as Tsai's*) in *re-projection accuracy* and *stability* on six clips of an mpeg2 video with resolution 704×576 recorded by a Panasonic DVD recorder from TV signal. More evaluations and applications of our algorithm are presented in another paper of ours [9].

3.1. Comparison on Re-projection Accuracy

We use the function defined in equation (10) to measure the re-projection similarity. We test our and Tsai's algorithms on six clips and the results are given in Table 1. Table 1 shows that our algorithm is above 10% better than Tsai's in re-projection similarity.

Table 1. Comparison on projection similarity between our algorithm and Tsai's algorithm.

clip	1	2	3	4	5	6
Ours	0.7533	0.726	0.7219	0.7354	0.7598	0.7661
Tsai's	0.6266	0.645	0.6453	0.5337	0.6733	0.5921
% Δ	16.819	11.166	10.611	27.427	11.385	22.712

3.2. Comparison on Stability of Camera Parameters

For broadcast tennis video (BTV), we assume that γ is constant and s=0 (see equations 7-8). Among the remaining parameters, our concern mainly are the camera center (*shorted as* CC) and focus length *f*. Table 2 gives the average of fluctuations of CC, *f*, (u_0,v_0) and angles of our and Tsai's algorithms (*fluctuation is the difference of two values of the same parameters and it is used to measure the stability of parameters [3]*). Table 2 shows that for the fluctuations of CC ours are at most 16.6% of Tsai's and that for the fluctuations of *f* ours are at most 18.2% of Tsai's. As for fluctuations of (u_0,v_0) and angles, no algorithm is uniformly better than the other. According to our experiments, Tsai's algorithm is very accurate as a general algorithm. However, our algorithm can outperform it for BTV since we use techniques of frame grouping and Hough-like search. It is important that Hough-like search can work with sophisticated measure function.

Table 2. Comparison on stablity b/w our and Tsai's algorithms.

	clip	**1**	**2**	**3**	**4**	**5**	**6**
CC	Tsai's	42.14	11.62	49.49	7.362	65.20	13.58
	Ours	2.964	0.904	0.463	0.174	0.807	2.25
	ratio	7.03%	7.78%	0.94%	2.36%	1.24%	16.6%
f	Tsai's	65.27	30.39	52.24	20.46	104.5	35.25
	Ours	5.562	4.009	2.554	0.796	5.169	6.427
	ratio	8.52%	13.2%	4.89%	3.89%	4.94%	18.2%
u_0 & v_0	Tsai's	2.113	2.351	4.719	1.464	2.040	2.169
	Ours	1.655	1.011	1.242	0.218	1.009	2.682
angles	Tsai's	0.171	0.013	0.039	0.008	0.171	0.017
	Ours	0.408	0.420	0.003	0.001	0.408	0.438

4. CONCLUSIONS AND FUTURE WORK

We have presented an original camera calibration algorithm which can acquire accuracy and stability camera parameters for broadcast tennis video (BTV). It uses two techniques: frame grouping and Hough-like search. The grouping technique helps to acquire more stable and accurate camera parameters with the by-product that the frames with erroneous features are singled out. A Hough-like search then tunes some parameters. Comparisons with Tsai's confirm the merits of our algorithm.

Two of many other future jobs remain to be done. For camera calibration, we want to extend our technique to other types of sports videos such as soccer, badminton, etc. For application, we want to use the results of camera calibration to aid the video analysis and video enhancement.

5. REFERENCES

[1] D. Farin, J. Han, and P. H. N. de With. Fast camera calibration for the analysis of sport sequences, *ICME*2005, pp482-485.

[2] A. Fusiello, A. Benedetti, M. Farenzena, and A. Busti. Globally convergent autocalibration using interval analysis, *PAMI*, 26(12): 1633-1638, 2004.

[3] J. I. González, J. C. Gámez, C. G. Artal and A. M. N. Cabrera. Stability study of camera calibration methods, *CI Workshop en Agentes Físicos*, Spain, WAF 2005.

[4] R. Hartley and A. Zisserman. Multiple view geometry in computer vision, *Cambridge U. Press*, 2003 (2nd edition), UK.

[5] E. E. Hemayed. A survey of camera self-calibration, *IEEE Conf. on Advanced Video & Signal Based Surveil.*, pp351-357, 2003.

[6] I. D. Reid and A. Zisserman. Goal-directed video metrology, *ECCV*1996, vol. II, pp647–658.

[7] G. Sudhir, J. C. M. Lee, and A. K. Jain. Automatic classification of tennis video for high-level content-based retrieval, IEEE Workshop on CBAIVD, pp81-90, 1998, in conj. with ICCV98.

[8] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the shelf TV cameras and lenses, *IEEE J. of Robotics and Automation*, vol. RA-3, No. 4, Aug 1987, pp 323-344.

[9] X. Yu, N. Jiang, L.-F. Cheong, H. W. Leong, and X. Yan. Automatic camera calibration of broadcast tennis video with applications to 3D virtual content insertion and ball detection and tracking, CVIU, to appear soon.

[10] Z. Zhang. A flexible new technique for camera calibration, *PAMI*, 22(11):1330-1334, 2000.