Accurate and Stable Camera Calibration of Broadcast Tennis Video

1Xinguo Yu, 2Nianjuan Jiang, and 3Loong-Fah Cheong

1Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, xinguo@i2r.a-star.edu.sg
2Dept of Electrical and Computer Engineering, National University of Singapore, Singapore 117543, {yu0308314, elecf1}@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° 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 $3 \times 4$ camera projection matrix.

Then for a pinhole camera, the mapping between the 3D world and the 2D image is written compactly as

$$m \equiv Pw$$  \hspace{1cm} (1)

where $\equiv$ means that two sides can differ by an arbitrary scale.
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 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 x in the lower part of the figure.

2.2. Ground-Feature Extraction

2.2.1. Initial Ground Feature Extraction

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 \( 1Y, (X, m) \).

\[ m \cong Hw. \quad (2) \]

Similarity Measure of Two Court Images: Let \( C_{\text{std}} \) and \( C_{\text{com}} \) 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 \( L_{i, \text{std}} \) and \( L_{i, \text{com}} \) to denote the sets of all the points contained in the straight line \( L_i \) in \( C_{\text{std}} \) and \( C_{\text{com}} \) respectively for \( i = 1 \) to 9. The measure function \( M_l(i) \) of straight line \( L_i \) is defined as

\[ M_l(i) = \frac{|L_{i, \text{std}}| \cdot |L_{i, \text{com}}|}{|L_{i, \text{std}}|}. \quad (3) \]

where \(|\cdot|\) is the cardinality of a set.

The measure function \( M_{\text{std}}(C_{\text{std}}, C_{\text{com}}) \) for measuring the similarity of the two courts is defined as:

\[ M_{\text{std}}(C_{\text{std}}, C_{\text{com}}) = \frac{1}{9} \sum_{i=1}^{9} \psi(i). \quad (4) \]

where \( \psi(i) = M_l(i) \) if \( M_l(i) > \beta(\text{threshold}) \); otherwise = 0.

Measure Function of Homography: Let \( C_{\text{ground}} \) be the ground model of the tennis court in the world coordinate system. We use \( C_{\text{hori}}(H, C_{\text{ground}}) \) to denote the transformed court from the ground model of the tennis court onto the image plane according to \( H \). Let \( C_{\text{img}} \) be the segmented court in frame F.
With the two court images $C_{\text{img}}$ and $C_{\text{hom}}$, we define the measure function $M_{l}(H, F)$ for a given $H$ and a frame $F$ as

$$M_{l}(H, F) = M_{\text{det}}(C_{\text{hom}}, C_{\text{img}}).$$  \hspace{1cm} (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 $M_{l}(H, F)$ 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 $V_{\text{max}} = M_{l}(H_0, F)$; $H_{td} = H_0$.

For all $H$ in $H_{sp}$ do

- If $M_{l}(H, F) > V_{\text{max}}$, then $V_{\text{max}} = M_{l}(H, F)$; $H_{td} = H$.
- 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 $(1)_{g1}, I_{g1}, K_{g1}) = H_{f} w_{g1}, x_{g1} = I_{g1} / K_{g1}, y_{g1} = J_{g1} / K_{g1} (i=1, 2)$ 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 $L_{0} = (x_{g1}, y_{g1})$ and $R_{0} = (x_{g2}, y_{g2})$ 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 $C_{\text{phy}}$ be the complete model of the tennis court that includes both the ground model and the two net poles. We use $C_{\text{phy}}(P, C_{\text{phy}})$ to denote the projected court from $C_{\text{phy}}$ via $P$ according to equation (1). $P$ will be computed based on $(C1, C2, ..., C14)$ $\cup$ $\{L, R\}$ (see next subsection). With the obtained $P$, we can formulate, for a given image $F$, the measure function $M_{s}(L, R, F)$ to tune the coordinates of the pole-tops.

$$M_{s}(L, R, F) = M_{\text{det}}(C_{\text{phy}}, C_{\text{img}}).$$  \hspace{1cm} (6)

Based on the initial coordinates of the pole-tops, we create a Hough space $H_{pol} = \{ (C1, C2, ..., C14) \cup (x_{G1}, y_{G1} + \tau) \cup (x_{G2}, y_{G2} + \tau), \text{with } \tau \text{ varying from -2 to 2 in a step size of 0.01} \}$ 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 $H_{pol}$

*Step 2:* Initialize $V_{\text{max}} = M_{s}(L_{0}, R_{0}, F)$; $L_{td} = L_{0}$ and $R_{td} = R_{0}$.

For all $L$ and $R$ in $H_{pol}$ do

- If $M_{s}(L, R, F) > V_{\text{max}}$, then $V_{\text{max}} = M_{s}(L, R, F)$; $L_{td} = L$ and $R_{td} = R$.
- Terminate and output $L_{td}$ and $R_{td}$.

**2.3.3. Camera Matrix Factorization**

From $m \equiv Pw$ (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

$$P = KR[I - C].$$  \hspace{1cm} (7)

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

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

where $f$ is focal length, $\gamma$ 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 $(\theta_{x}, \theta_{y}, \theta_{z})$. These parameters form the group-invariant parameters. The remaining two parameters, namely the aspect ratio $\gamma$ 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. More exactly, the homography from the image space to the 3D world space is given by:

$$w = H_{pm}m$$  \hspace{1cm} (9)

where $H_{pm} = H^{-1}$. Let $m_{pm} = (w/2, h/2, 1)$ where $(w/2, h/2)$ is the image center. Then $H_{pm} m_{pm}$ 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 $\gamma$ is constant in a clip, we find $\gamma$ 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

$$M_{3}(P, F) = \frac{1}{11} \sum_{i=1}^{11} \psi(i)$$  \hspace{1cm} (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 \( H_{\text{base}} \) 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 \( H_{\text{base}} \) and choose the best camera projection matrix.

Procedure 3

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

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 704x576 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.

<table>
<thead>
<tr>
<th>clip</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.7533</td>
<td>0.726</td>
<td>0.7219</td>
<td>0.7354</td>
<td>0.7598</td>
<td>0.7661</td>
</tr>
<tr>
<td>Tsai’s</td>
<td>0.6266</td>
<td>0.645</td>
<td>0.6453</td>
<td>0.5337</td>
<td>0.6733</td>
<td>0.5921</td>
</tr>
<tr>
<td>% Δ</td>
<td>16.819</td>
<td>11.166</td>
<td>10.611</td>
<td>27.427</td>
<td>11.385</td>
<td>22.712</td>
</tr>
</tbody>
</table>

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, \( f \), \( u_0 \) and \( v_0 \) 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 \( u_0 \) and \( v_0 \) 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.

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