Video-based sequential Bayesian homography estimation for soccer field registration

Paul Johannes Claasen *, Johan Pieter de Villiers

Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Lynnwood Road, Hatfield, Pretoria, 0028, South Africa

A R T I C L E   I N F O

A B S T R A C T

A novel Bayesian framework is proposed, which explicitly relates the homography of one video frame to the next through an affine transformation while explicitly modelling keypoint uncertainty. The literature has previously used differential homography between subsequent frames, but not in a Bayesian setting. In cases where Bayesian methods have been applied, camera motion is not adequately modelled, and keypoints are treated as deterministic. The proposed method, Bayesian Homography Inference from Tracked Keypoints (BHITK), employs a two-stage Kalman filter and significantly improves existing methods. Existing keypoint detection methods may be easily augmented with BHITK. It enables less sophisticated and less computationally expensive methods to outperform the state-of-the-art approaches in most homography evaluation metrics. Furthermore, the homography annotations of the WorldCup and TS-WorldCup datasets have been refined using a custom homography annotation tool that has been released for public use. The refined datasets are consolidated and released as the consolidated and refined WorldCup (CARWC) dataset.

1. Introduction

Homography estimation generally refers to a planar projective transformation that relates corresponding points in two views of the same scene. Homography estimation plays a crucial role in many computer vision applications. Examples include automated panoramic image stitching (Brown & Lowe, 2007), simultaneous localisation and mapping (SLAM) (DeTone et al., 2016; Le et al., 2020; Mur-Artal et al., 2015), camera calibration and pose estimation (Citraro et al., 2020; Zhang, 1999) and video stabilisation (Vlahović et al., 2018). Sports field registration applies specifically to the case where one of the scene views represents a structured model of a sports field. Sports field registration enables aligning virtual overlays, such as graphics, annotations, or analysis tools, with the real-world sports field, as well as accurate player tracking and augmented reality experiences.

While the application of interest in this work is sports field registration, the Bayesian homography estimation method presented in this paper could also be applied to some of these other applications under certain conditions. For example, given a suitable keypoint matching algorithm and reasonable estimates of the process and measurement noise parameters, the method could be applied to estimate the homography of planar scenes in SLAM applications (Mur-Artal et al., 2015). It may also be employed to estimate camera pose as in Citraro et al. (2020). While this paper focuses on the specific use case of soccer

field registration, it is hoped to inspire similar approaches in these application areas.

1.1. Contributions

The main contributions of this paper are as follows:

- Whereas previous methods treat keypoints used for homography estimation as deterministic, this work considers them stochastic.
- A novel dynamics model is proposed which explicitly relates the homography of immediately subsequent frames to one another.
- A two-stage Kalman filter approach is used to filter the homography from tracked keypoints.
- The proposed approach improves the results of a relatively simple keypoint detector to outperform an expensive state-of-the-art method.
- Finally, the WorldCup (Homayounfar et al., 2017) and TS-WorldCup (Chu et al., 2022) datasets have been refined using a custom homography annotation tool (released for public use) and released as the consolidated and refined WorldCup (CARWC) dataset.

The remainder of the article is structured as follows. Section 2 provides the necessary background of projective spaces to understand what

---

the homography matrix represents. Section 3 provides an overview of previous works in the literature concerning homography estimation. Section 4 describes the proposed approach in detail, including mathematical derivations, and summarises the main differences between the proposed approach and previous methods from the literature. Section 5 describes the experiments used to evaluate the proposed method. This section includes parameter settings, details regarding datasets, a description of how covariance matrices were measured and examples of their values and formulae for the evaluation metrics that were used. The results of these evaluations are presented and discussed in Section 6. Finally, the main conclusions are presented in Section 7.

2. Background

The world in $\mathbb{R}^3$ is imaged through a projective camera, resulting in a 3D projective space $\mathbb{P}^3$, which augments $\mathbb{R}^3$ with points at infinity (Hartley & Zisserman, 2004). A coordinate $\mathbf{X} = \begin{bmatrix} X & Y & Z \end{bmatrix}$ in $\mathbb{R}^3$ is augmented to form $\mathbf{X}' = \begin{bmatrix} X & Y & Z & T \end{bmatrix}^\top$ (termed a homogeneous coordinate) in $\mathbb{P}^3$, with the corresponding point at infinity occurring for $T = 0$. Although the 3D vector is augmented by the element T, the projective space is, by convention, still considered three-dimensional. Hence, the superscript of $\mathbb{P}$ remains 3. A projective camera then performs a linear mapping on the homogeneous coordinate $\mathbf{X}'$ from the 3D projective space $\mathbb{P}^3$ (which represents the world space) to the homogeneous coordinate $\mathbf{x}' = \begin{bmatrix} x & y & t \end{bmatrix}^\top$ in the 2D projective space $\mathbb{P}^2$ (which represents the image space). The transformation from 3D to 2D projective space is governed by

$$\begin{bmatrix} x \\ y \\ t \end{bmatrix} = \mathbf{P} \begin{bmatrix} X \\ Y \\ Z \\ T \end{bmatrix},$$

where $\mathbf{P} = \kappa\,[\mathbf{R}|\mathbf{C}]$ and | denotes column-appended matrix augmentation. $\kappa \in \mathbb{R}^{3\times3}$ represents the internal camera parameters and has the general form

$$\kappa = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix},$$

where α_x and α_y are scale factors in the x- and y-coordinate directions, respectively, s is the skew which is non-zero if the x- and y-axes are not perpendicular and $\begin{bmatrix} x_0 & y_0 \end{bmatrix}^\top$ represents the coordinates of the principal point — the geometric centre of the image. For further details regarding α_x, α_y, s, x_0 and y_0, the reader is referred to Hartley and Zisserman (2004). $\mathbf{R} \in \mathbb{R}^{3\times3}$ and $\mathbf{C} \in \mathbb{R}^3$ respectively relate the camera orientation (rotation) and position (translation) to the world coordinate system. These are the external camera parameters (Hartley & Zisserman, 2004). An arbitrary homogeneous vector $\mathbf{x}' = \begin{bmatrix} x & y & t \end{bmatrix}^\top$ in $\mathbb{P}^2$ may be normalised to become $\begin{bmatrix} x/t & y/t & 1 \end{bmatrix}^\top$, $t \neq 0$, which represents

$$\begin{bmatrix} x/t & y/t \end{bmatrix}^\top \tag{1}$$

in $\mathbb{R}^2$, a point in the image (Hartley & Zisserman, 2004). That is, for a constant $\mathbf{x}'$, $k\mathbf{x}'$ represents the same point in the image for any $k \neq 0$ and may be thought of as a ray in $\mathbb{R}^3$ passing through the centre of projection of the camera. One may consider the projective space $\mathbb{P}^2$ as a space consisting of the set of such rays, each representing a single point in the image in $\mathbb{R}^2$. Finally, if all the world points are coplanar such that $Z = 0$, the transformation from 3D to 2D projective space is performed by the homography matrix $\mathbf{H}$:

$$\begin{bmatrix} x \\ y \\ t \end{bmatrix} = \mathbf{H} \begin{bmatrix} X \\ Y \\ T \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} X \\ Y \\ T \end{bmatrix}. \tag{2}$$

It should be noted that $\mathbf{H}$ is determined up to a scale and thus has only eight degrees of freedom (Hartley & Zisserman, 2004).

3. Related work

This section examines previous homography estimation methods from the literature to illustrate the novelty of the proposed method.

3.1. Homography estimation

Homography estimation is typically achieved by identifying corresponding features, or keypoints, between two images. This may be achieved through the use of features extracted by methods such as the scale-invariant feature transform (SIFT) (Lowe, 2004) or ORB (Oriented FAST and Rotated BRIEF, where FAST refers to a keypoint detection method and BRIEF refers to another feature descriptor (Rublee et al., 2011)), and matching methods such as k-nearest neighbours (Brown & Lowe, 2007) or bags of words (Mur-Artal et al., 2015). These keypoints are subsequently used to estimate the mapping between the images with the direct linear transform (DLT) (Hartley & Zisserman, 2004) or random sample consensus (RANSAC) (Fischler & Bolles, 1987) algorithms. Another approach to homography estimation is that of iterative optimisation such that the alignment between a target image and the transformation of another image is maximised through the minimisation of a chosen loss function (Evangelidis & Psarakis, 2008; Lucas & Kanade, 1981). These methods are usually slower than feature-based methods. Still, the robustness and accuracy of feature-based methods are subject to the number of detected keypoints and the accuracy with which keypoint correspondences can be determined (Hartley & Zisserman, 2004). Thus, feature-based methods may be less reliable when few correspondences exist due to large view differences or where the extracted features are not sufficiently salient due to image-specific lighting or noise. Methods that rely on estimating a differential homography between subsequent video frames and other features have been proposed (Nishida et al., 2011; Simon et al., 2000). Indeed, the authors note that their methods fail under some lighting conditions. Recent methods leverage deep neural networks to regress a parameterisation of the homography matrix directly (DeTone et al., 2016; Le et al., 2020; Nguyen et al., 2018; Vlahović et al., 2018), using features directly computed from a tuple of image patches for which a homography estimate is desired.

3.2. Sports field registration

Various recent methods of sports field registration identify corresponding features between the field template and camera image with the use of deep (convolutional) neural networks (Chen & Little, 2018; Chu et al., 2022; Citraro et al., 2020; Cuevas et al., 2022; Homayounfar et al., 2017; Nie et al., 2021; Sha et al., 2020; Theiner & Ewerth, 2022). These are subsequently used to estimate a mapping between the image and field template. This estimation may be achieved with the DLT (Hartley & Zisserman, 2004) or RANSAC (Fischler & Bolles, 1987) algorithms (Chu et al., 2022; Citraro et al., 2020; Nie et al., 2021; Theiner & Ewerth, 2022). Some methods refine this estimate with, or otherwise rely entirely on, a combination of the following: regressing the homography directly from the input image and field model with a deep neural network (Jiang et al., 2019; Shi et al., 2022), obtaining an estimate for the camera pose by matching with a feature-pose database (Chen & Little, 2018; Sha et al., 2020), iterative optimisation of the camera pose or homography based on re-projection error or some other metric (Chen & Little, 2018; Jiang et al., 2019; Sha et al., 2020; Shi et al., 2022; Theiner & Ewerth, 2022), or the use of a Markov Random Field (MRF) (Homayounfar et al., 2017).

Using a feature-pose database is cumbersome and will almost always require additional optimisation. Furthermore, methods that rely on optimisation tend to be slow, from experiments with Theiner and

Ewerth (2022). Similarly, the authors of Homayounfar et al. (2017) report many average iterations and an average inference time of 0.44 s with their MRF-based method when applied to soccer field registration.

In comparison, keypoint-detection-based models are attractive due to their relatively small computational footprint and the potential to use the detected keypoints in a Bayesian framework, which is the approach in this paper.

3.3. Tracking

Few existing methods exploit the temporal consistency between subsequent video frames, with some exceptions being found in Citraro et al. (2020), Nie et al. (2021), Nishida et al. (2011), Simon et al. (2000). The differential homography between video frames is used in Nishida et al. (2011), Simon et al. (2000) for road plane detection (which is important in autonomous driving applications) with optical flow and human-assisted keypoint-less tracking, respectively. However, neither of these methods considers the differential homography in a Bayesian setting. Nie et al. (2021) make use of online homography refinement by minimising two loss functions, one of which also takes into account the relative homography $\hat{\mathbf{H}}_t \hat{\mathbf{H}}_{t-1}^{-1}$ between the current and previous frame as in Nishida et al. (2011), Simon et al. (2000). Once again, this is not performed with a Bayesian treatment of the homography. Finally, (Citraro et al., 2020) makes use of player positions and field keypoints detected by a U-Net architecture (Ronneberger et al., 2015). A homography estimate is obtained for each frame in a sequence and decomposed to estimate camera intrinsic and extrinsic parameters. A condensation particle filter (Isard & Blake, 1998) is applied, which only considers and enforces temporal consistency on the external camera parameters i.e. the dynamics model applied is $[\mathbf{R} \mid \mathbf{C}]_t = [\mathbf{R} \mid \mathbf{C}]_{t-1} + \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with particle weights obtained by a re-projection metric. While this method employs the Bayesian particle filter framework, the dynamics model is unsuitable. Particularly, camera movement is not effectively modelled: changes in the estimated pose are modelled as noise. Additionally, keypoint uncertainty is not taken into account. Indeed, a heuristic measure is required to determine when the filter should be re-initialised after it inevitably diverges.

To the best of the authors' knowledge, the literature has not explored a Bayesian approach that explicitly incorporates homography, field template and keypoint measurement uncertainty while also modelling relative camera motion.

4. Approach

The proposed approach, Bayesian Homography Inference from Tracked Keypoints (BHITK), is inspired by recent developments in tracking-by-detection methods. Specifically, those which employ a form of camera motion compensation. It has been shown that tracking performance can be improved by transforming bounding boxes forecasted at time $t − 1$ such that they align more closely with detections at time t. This transformation effectively estimates and corrects for the non-stationarity in measurements induced by camera motion. This is performed in Bergmann et al. (2019), Han et al. (2020), Khurana et al. (2020) with the image registration algorithm in Evangelidis and Psarakis (2008), which estimates a non-linear mapping of pixels from one frame to the next. Another method makes use of ORB (Rublee et al., 2011) and RANSAC (Fischler & Bolles, 1987) to quickly align subsequent frames (Du et al., 2021). The global motion compensation technique of the Video Stabilisation module of OpenCV (Bradski, 2000) is used instead in Aharon et al. (2022). Its use is motivated by sparse optical flow features and translation-based local outlier rejection, which allows the resulting affine matrix estimated by RANSAC to be largely unaffected by dynamic objects. Since this method is more focused on background motion, it is ideal for use in the present case where several dynamic objects (e.g. soccer players, the ball, referees and spectator movement) are expected to be present.

Table 1
A comparison between the approaches of BHITK and related methods.

Method	Relative (inter-frame) homography or pose	Bayesian framework	Keypoint uncertainty (fully Bayesian approach)[a]
Nishida et al. (2011)	✓	X	X
Simon et al. (2000)	✓	X	X
Nie et al. (2021)	✓	X	X
Citraro et al. (2020)	Modelled inappropriately	✓	X
BHITK (proposed)	✓	✓	✓

[a] Equivalently, whether RANSAC is used to obtain the homography from noisy point estimates (indicated with a cross) or whether the homography is inferred with the fusion of different keypoint distributions (indicated with a checkmark).

As noted by Nie et al. (2021), sparse features due to generally texture-less sports fields, narrow camera field of view, and occlusion by players represent the most significant challenges to keypoint-detection-based methods since these challenges lead to fewer detected keypoints and consequently a less robust homography estimate. This work proposes a dynamics model that explicitly relates image keypoint positions from one frame to the next. A relation between subsequent homographies is derived from this, appropriately considering camera motion. A Kalman filter framework with linear and non-linear components is used, encompassing the homography as part of the state vector. Thus, even when few or no keypoints are detected in narrow-field-of-view or occluded situations, the homography estimate forecasted by the dynamics model – which is independent of specific field template keypoints – may still be reasonably accurate due to the incorporation of the history of keypoints that were visible up until that time, and the possible estimation of out-of-frame keypoint positions. Furthermore, keypoint noise is modelled explicitly, contributing to even more accurate homography estimation. Contrary to the status quo, RANSAC (Fischler & Bolles, 1987) is only used to obtain an initial homography estimate. Thereafter, the homography is inferred solely from the dynamics and measurement models and the fusion of the measured keypoint statistics. Whereas RANSAC considers keypoint measurements to be noisy point estimates, the proposed method instead considers the entire estimated distribution of each keypoint. The approach is flexible and can extend existing keypoint detection methods. Table 1 summarises the contributions of BHITK compared to related methods in the literature.

4.1. Derivations

Given a set of N known field template keypoints represented by normalised homogeneous world coordinates $\{\mathbf{X}^{F,j} \in \mathbb{P}^2 | 1 \leq j \leq N\}$, and a set of N image keypoints represented by normalised homogeneous image coordinates $\{\mathbf{x}^{I,j} \in \mathbb{P}^2 | 1 \leq j \leq N\}$, the corresponding points in these sets are assumed to be coplanar in world coordinates. The goal is to estimate the homography $\mathbf{H}$, which relates keypoints in the image and their corresponding coordinates in the field template. Assume that an image motion $\mathbf{A}$ is available at each time step t such that

$$\mathbf{x}_t^{I,j} = \mathbf{A}_t \mathbf{x}_{t-1}^{I,j}. \tag{3}$$

Furthermore, since field template keypoints are constant:

$$\mathbf{X}_t^{F,j} = \mathbf{X}_{t-1}^{F,j}, \tag{4}$$

where the dependence on time is retained since time-dependent random samples are later added to model the possible uncertainty of field template keypoint positions. From (2), the coordinates in $\mathbb{P}^2$ of an
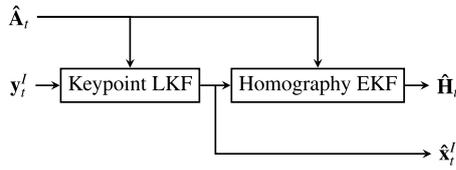
Fig. 1. Implemented Kalman filter framework. The first stage filters measured keypoint positions $\mathbf{y}_t^I$ according to the estimated affine transformation $\hat{\mathbf{A}}_t$. The EKF makes use of the filtered positions, $\hat{\mathbf{x}}_t^I$, and the estimated affine transformation to infer an estimate of the homography, $\hat{\mathbf{H}}_t$.

image keypoint $\mathbf{x}^{I,j}$ which corresponds to a field template keypoint $\mathbf{X}^{F,j}$ may be obtained by

$$\mathbf{x}_t^{I,j} = \mathbf{H}_t \mathbf{X}_t^{F,j}. \tag{5}$$

Substituting (5) into (3):

$$\mathbf{x}_t^{I,j} = \mathbf{A}_t \mathbf{H}_{t-1} \mathbf{X}_{t-1}^{F,j}.$$

Making use of the relation in (4):

$$\mathbf{x}_t^{I,j} = \mathbf{A}_t \mathbf{H}_{t-1} \mathbf{X}_t^{F,j}.$$

Finally, comparing this result with (5):

$$\mathbf{H}_t = \mathbf{A}_t \mathbf{H}_{t-1}. \tag{6}$$

In practice, it is necessary to obtain the image coordinates in $\mathbb{R}^2$ represented by $\mathbf{x}_t^{I,j} \in \mathbb{P}^2$. This is achieved by normalising the homogeneous $\mathbf{x}_t^{I,j}$ with respect to its last element, as illustrated in (1). Let $\mathrm{norm}(\cdot)$ denote this normalisation, such that $\mathrm{norm}(\begin{bmatrix} x & y & z \end{bmatrix}^\top) = \begin{bmatrix} x/z & y/z & 1 \end{bmatrix}^\top$. Thus, (5) becomes

$$\mathbf{x}_t^{I,j} = \mathrm{norm}\left(\mathbf{H}_t \mathbf{X}_t^{F,j}\right). \tag{7}$$

$\mathbf{A}_t$ is an affine transformation matrix which allows for translation, rotation and uniform scaling in the x and y image dimensions. It is estimated at each time step with the global motion compensation method of OpenCV (Bradski, 2000):

$$\hat{\mathbf{A}}_t = \left[\begin{array}{c|c} \mathbf{A}_t^u \in \mathbb{R}^{2\times2} & \mathbf{b}_t \in \mathbb{R}^{2\times1} \\ \hline \mathbf{0}^{1\times2} & 1 \end{array} \right],$$

where $\mathbf{A}_t^u$ is a rotation and scaling matrix and $\mathbf{b}_t$ is a translation vector, such that, for arbitrary vectors, $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$,

$$\mathbf{y} = \mathbf{A}_t^u \mathbf{x} + \mathbf{b}_t$$

is equivalent to

$$\begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} = \hat{\mathbf{A}}_t \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

in homogeneous coordinates. It is easily shown that, for an arbitrary $\mathbf{x} \in \mathbb{P}^2$, $\mathrm{norm}(\mathbf{A}_t \mathbf{x}) = \mathbf{A}_t \mathrm{norm}(\mathbf{x})$. Therefore, the relation obtained in (6) remains valid despite the alteration to (5) given in (7). Note that this is not generally the case if $\mathbf{A}_t$ is replaced with, e.g. an inter-frame homography (i.e. a homography that provides a mapping between the current and previous frames) where the last row is not $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$, which would require additional normalisation. Therefore, the proposed affine transformation enables the relationship between subsequent homographies to be expressed and precludes normalisation in (3), maintaining linearity in the keypoint dynamics model. Another advantage of using the proposed affine transformation is that it is independent of detecting specific, pre-defined keypoints. Thus, the absence of any number of such keypoints is assumed not to affect the dynamics model significantly. Nevertheless, incorporating a robust, global transformation of keypoints between subsequent frames (which may also be

non-linear) into the state vector is a compelling prospect but deemed a topic for future research. For now, it is assumed that uncertainties in the estimation of $\mathbf{A}_t$ are largely mitigated by modelling (3) and (6) as stochastic processes. Uncertainty in field template keypoint positions may also be specified, although this inclusion may only benefit practical applications where field dimensions vary. The dynamics model is thus obtained by treating (3), (4) and (6) as random variables:

$$\mathbf{x}_t^{I,j} = \mathbf{A}_t \mathbf{x}_{t-1}^{I,j} + \mathbf{w}_t^{I,j}, \tag{8}$$

$$\mathbf{X}_t^{F,j} = \mathbf{X}_{t-1}^{F,j} + \mathbf{w}_t^{F,j}, \tag{9}$$

$$\mathbf{H}_t = \mathbf{A}_t \mathbf{H}_{t-1} + \mathbf{W}_t^H, \tag{10}$$

where $\mathbf{w}_t^{I,j} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{I,j}\right)$ and $\mathbf{w}_t^{F,j} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{F,j}\right)$. The elements of $\mathbf{W}_t^H$ are drawn from $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^H \in \mathbb{R}^{9\times9}\right)$. Similarly, the measurement model is obtained from (7):

$$\mathbf{x}_t^{I,j} = \mathrm{norm}\left(\mathbf{H}_t \mathbf{X}_t^{F,j}\right) + \mathbf{w}_t^{M,j}, \tag{11}$$

where $\mathbf{w}_t^{M,j} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{M,j})$. The last element of each homogeneous coordinate $\mathbf{x}_t^{I,j}$ and $\mathbf{X}_t^{F,j}$ is always 1. For the remainder of this paper, these coordinates are assumed to be transformed to $\mathbb{R}^2$ by simply omitting their last elements. Therefore, the distributions from which $\mathbf{w}_t^{I,j}$, $\mathbf{w}_t^{F,j}$ and $\mathbf{w}_t^{M,j}$ are drawn are also considered elements of $\mathbb{R}^2$, with corresponding covariance matrices in $\mathbb{R}^{2\times2}$.

The above dynamics and measurement models are used in a two-stage Kalman filter. Fig. 1 concisely illustrates the roles of the estimated affine transformation $\hat{\mathbf{A}}_t$, keypoint measurements $\mathbf{y}_t^I$ and two-stage filter in the proposed approach. The filter stages are subsequently described in detail.

4.2. Linear keypoint filter

The left-hand side of (8)–(10) represents the state space under consideration. The state elements are not independent. A given set of keypoint measurements may be related to the state in two ways: either directly since the keypoint positions are part of the state vector or through the re-projection of the field keypoints to image keypoints by the homography. Although not strictly required for homography inference, image keypoint positions are retained as part of the state vector. This can improve the homography estimation since the filtered keypoint positions are likely more accurate, as long as the zero-mean Gaussian measurement noise assumption is reasonable and the process and measurement covariances are appropriately tuned. Hence, a two-stage approach is proposed. The first stage consists of a linear Kalman filter (LKF), which considers the image keypoints the only part of its state vector. The state vector takes the form $\mathbf{x}_t^I = \left[\left(\mathbf{x}^{I,1}\right)^\top \cdots \left(\mathbf{x}^{I,j}\right)^\top \right]_t^\top$, where $\mathbf{x}^{I,j} \in \mathbb{R}^2, j \leq N$. Its dynamics are governed by (8). The prediction step is implemented as follows:

$$\hat{\mathbf{x}}_{t|t-1}^I = \tilde{\mathbf{A}}_t \hat{\mathbf{x}}_{t-1|t-1}^I + \tilde{\mathbf{B}}_t,$$

$$\mathbf{P}_{t|t-1}^I = \tilde{\mathbf{A}}_t \mathbf{P}_{t-1|t-1}^I \tilde{\mathbf{A}}_t^\top + \mathbf{Q}^I,$$

where $\hat{\mathbf{x}}_{t|t-1}^I$ and $\hat{\mathbf{x}}_{t-1|t-1}^I$ denote the estimates of the means of the predicted and filtered states at time $t-1$, respectively. Similarly, $\mathbf{P}_{t|t-1}^I$ and $\mathbf{P}_{t-1|t-1}^I$ denote the predicted and filtered estimates of the state covariance matrix. $\mathbf{Q}^I$ represents the process noise covariance matrix. Finally, similar to Aharon et al. (2022):

$$\tilde{\mathbf{A}}_t = \begin{bmatrix} \mathbf{A}_t^u & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_t^u \end{bmatrix}, \tilde{\mathbf{B}}_t = \begin{bmatrix} \mathbf{b}_t \\ \vdots \\ \mathbf{b}_t \end{bmatrix}.$$

Upon receiving K keypoint measurements $\mathbf{y}_t^I = \left[\left(\mathbf{y}_1^{I,j}\right)^\top \cdots \left(\mathbf{y}_k^{I,j}\right)^\top \right]_t^\top$, where $\mathbf{y}_k^{I,j} \in \mathbb{R}^2, k \leq K$, the update step is performed:

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}^I \mathbf{H}_t^\top \left(\mathbf{H}_t \mathbf{P}_{t|t-1}^I \mathbf{H}_t^\top + \mathbf{R}^I \right)^{-1},$$

$$\hat{\mathbf{x}}_{t|t}^I = \hat{\mathbf{x}}_{t|t-1}^I + \mathbf{K}_t \left(\mathbf{y}_t^I - \mathbf{H}_t \hat{\mathbf{x}}_{t|t-1}^I \right),$$

$$\mathbf{P}_{t|t}^I = \left(\mathbf{I} - \mathbf{K}_t \mathbf{H}_t \right) \mathbf{P}_{t|t-1}^I,$$

where $\mathbf{R}^I$ is the measurement noise covariance matrix and $\mathbf{H}_t \in \mathbb{R}^{2K \times 2N}$ is a matrix which consists of sub-matrices $\mathbf{h}_{k,j} \in \mathbb{R}^{2 \times 2}$ that relate the state to the measurements:

$$\mathbf{h}_{k,j} = \begin{cases} \mathbf{I}, & \text{if } \exists \mathbf{y}_k^{I,j} \in \mathbf{y}_t^I, \\ \mathbf{0}, & \text{otherwise}, \end{cases}$$

where $\mathbf{I}$ is the identity matrix.

4.3. Non-linear homography filter

The second stage of the proposed method directly incorporates the homography, in addition to the field template keypoints, into its state vector $\mathbf{x}_t^{FH} = \left[\left(\mathbf{X}^{F,1} \right)^\top \cdots \left(\mathbf{X}^{F,j} \right)^\top \quad \mathbf{h}_1^\top \quad \mathbf{h}_2^\top \quad \mathbf{h}_3^\top \right]_t^\top$, where $\mathbf{X}^{F,j} \in \mathbb{R}^2, j \leq N$ and $\mathbf{h}_1, \mathbf{h}_2$ and $\mathbf{h}_3$ respectively denote the first, second and third columns of the homography matrix. Its dynamics are governed by (9) and (10), which are linear processes. The prediction step is thus performed by

$$\hat{\mathbf{x}}_{t|t-1}^{FH} = \tilde{\mathbf{M}}_t \hat{\mathbf{x}}_{t-1|t-1}^{FH},$$

$$\mathbf{P}_{t|t-1}^{FH} = \tilde{\mathbf{M}}_t \mathbf{P}_{t-1|t-1}^{FH} \tilde{\mathbf{M}}_t^\top + \mathbf{Q}^{FH},$$

where $\mathbf{P}^{FH}$ and $\mathbf{Q}^{FH}$ denote the state and process covariance matrices, respectively. Furthermore,

$$\tilde{\mathbf{M}}_t = \begin{bmatrix} \mathbf{I}^{2N \times 2N} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{A}}_t \end{bmatrix}.$$

The non-linear measurement model (11) requires the transformation of the state by (7). This may be performed whilst retaining relatively high-order moments using the Unscented Transform in the Unscented Kalman Filter (UKF) (Wan & Merwe, 2000). However, since the relative homography between frames is expected to be small, the Extended Kalman Filter (EKF) approach is used instead by linearising around the current state estimate. Enforcing this belief in the UKF requires tuning its hyper-parameters (i.e. how close to the mean sigma points are sampled), which is avoided. Using the UKF slightly degraded performance due to significant errors before convergence. The update step is therefore performed as follows:

$$\mathbf{K}_t^{FH} = \mathbf{P}_{t|t-1}^{FH} \mathbf{J}_t^\top \left(\mathbf{J}_t \mathbf{P}_{t|t-1}^{FH} \mathbf{J}_t^\top + \mathbf{P}_{t|t}^I \right)^{-1},$$

$$\hat{\mathbf{x}}_{t|t}^{FH} = \hat{\mathbf{x}}_{t|t-1}^{FH} + \mathbf{K}_t^{FH} \left(\hat{\mathbf{x}}_{t|t}^I - \mathrm{h} \left(\hat{\mathbf{x}}_{t|t-1}^{FH} \right) \right),$$

$$\mathbf{P}_{t|t}^{FH} = \left(\mathbf{I} - \mathbf{K}_t^{FH} \mathbf{J}_t \right) \mathbf{P}_{t|t-1}^{FH},$$

where $\mathrm{h}(\cdot)$ represents (7), $\hat{\mathbf{x}}_{t|t-1}^{FH}$ is augmented with a one in (7) to transform it to $\mathbb{P}^2$, and $\mathbf{J}_t$ is the Jacobian of $\mathrm{h}\left(\hat{\mathbf{x}}_{t|t-1}^{FH} \right)$ with respect to each of the state elements.

The proposed approach is adaptable to any keypoint detection method. Furthermore, the state and measurement models may be expanded to incorporate image distortion parameters. However, modelling distortion with a single-parameter division model as in Wu et al. (2017) slightly degraded performance and is therefore not considered, although such parameters may be useful in some practical applications.

5. Experiments

The following section describes the experiments used to evaluate BHITK in detail. This includes details on the practical implementation, parameter settings, datasets and evaluation metrics.

5.1. Practical considerations

The state mean estimates $\hat{\mathbf{x}}_{0|0}^I$ and $\hat{\mathbf{x}}_{0|0}^{FH}$ are initialised with the first measured keypoint positions, known field template positions and the initial homography estimate obtained with RANSAC. Since the homography matrix is determined up to a scale (2), the initial estimate is normalised with respect to h_{33}. The Kalman filter state vector excludes this element (h_{33}).

For the current purposes, $\mathbf{\Sigma}^{F,j} = \mathbf{0} \forall j$. The other process covariance matrices $\mathbf{\Sigma}^{I,j}$ and $\mathbf{\Sigma}^H$ are estimated empirically from the training data using the estimated affine transformation between training video frames, the ground truth keypoint positions and the ground truth homography annotations. Specifically, the mean-squared error (MSE) is used to estimate the variance of $\mathbf{\Sigma}^{I,j}$ and $\mathbf{\Sigma}^H$ in each state dimension, and the mean of the product of the errors of different state dimensions are used to estimate the covariances.

The measurement covariance matrix $\mathbf{\Sigma}^{M,j}$ of each keypoint j is estimated similarly using the measured and ground truth keypoint positions from the training set. The measured keypoint positions are used with RANSAC to produce a homography estimate for each training frame. These estimates are used with the ground truth homography annotations to estimate the covariance matrix with which the initial homography estimate is initialised. These are the only covariance matrices dependent on the keypoint detection method.

The process covariance matrices $\mathbf{Q}^I$ and $\mathbf{Q}^{FH}$, and the measurement covariance matrix $\mathbf{R}^I$ are constructed by concatenating the applicable covariance matrices obtained in the manner described above diagonally. The estimation of the covariance between distinct keypoints is complicated by the fact that keypoints do not always co-occur in the same image. Thus, independence between distinct keypoints is assumed. Finally, independence between distinct field template keypoints and between field template keypoints and the homography is also assumed. This follows from treating the field template positions as known ($\mathbf{\Sigma}^{F,j} = \mathbf{0} \forall j$), which is justified in the present case since the ground truth homography annotations are also obtained with this assumption.

While the discussion of the LKF and EKF in 4.2 and 4.3 imply that all of the known field template keypoint positions have corresponding image keypoints in $\mathbf{x}_t^I$, only the keypoints which have been measured at times prior to and including the current time step t are used in the EKF update step. Furthermore, the best results have been obtained when only the keypoints measured at the current time step t are used in the EKF update step. However, in the case of sparse keypoint positions, it may be helpful to initialise all image keypoint positions through the initial homography estimate and use all of the estimated keypoint positions in the EKF update, especially since the covariance estimates of those keypoints that have not been measured recently would be larger than that of those that have. The EKF takes this uncertainty into account.

5.1.1. Assumptions

The following summarises the major assumptions of the proposed BHITK method:

- The true playing field dimensions correspond to the template field dimensions.
- Eqs. (8)–(10) are reasonable approximations of the true state dynamics i.e. nonlinearities in the true state dynamics are negligible.
- Eq. (11) is a reasonable approximation of the true observation model i.e. lens distortion effects are negligible.
- The process and measurements noise terms in (8)–(11) are reasonably approximated by zero-mean Gaussian distributions.
- The linearisation error of the EKF is smaller than the measurement noise.

- Distinct keypoints are independent of one another in the image and ground plane, and the homography is independent of specific field template positions (see the discussion in 5.1).

5.2. Datasets

5.2.1. WC14 dataset

The WorldCup (WC14) dataset is typically used to evaluate soccer field registration (Chen & Little, 2018; Chu et al., 2022; Citraro et al., 2020; Homayounfar et al., 2017; Jiang et al., 2019; Nie et al., 2021; Sha et al., 2020; Shi et al., 2022; Theiner & Ewerth, 2022). It consists of 209 image-homography pairs in a training set and 186 in a test set. The images were obtained from broadcast television videos of the 2014 FIFA World Cup. The ground truth homography matrices are labelled manually. Unfortunately, as already noted (Theiner & Ewerth, 2022), the annotation of homography matrices is biased since the entire field is usually not visible in any given image. This problem is exacerbated by using too few ground-truth keypoints, i.e. a sparse keypoint annotation template focusing only on certain parts of the field. This is illustrated in Fig. 2(a), which shows an example of low-quality homography annotation in the WC14 dataset by re-projecting the grass band keypoints proposed in Cuevas et al. (2022) with the annotated homography. Notice that the keypoints do not align well with the grass bands, particularly those further away from the penalty area. Inadequate homography annotations undermine the reliability of the Intersection over Union (IoU) metrics often used to evaluate soccer field registration methods (Citraro et al., 2020; Theiner & Ewerth, 2022).

5.2.2. TS-WorldCup dataset

The TS-WorldCup dataset (TSWC) was introduced to augment the WC14 dataset since the WC14 dataset is relatively small (Chu et al., 2022). Unlike the WC14 dataset, the TSWC dataset consists of consecutive frames from 43 2014 and 2018 Soccer World Cup event videos. It contains 2925 and 887 images in its training and test sets. Similar to the WC14 dataset, the TSWC dataset suffers from annotation bias, albeit somewhat. This is illustrated in Fig. 3(a), where the annotation error is especially visible towards the bottom of the image.

5.2.3. CARWC dataset

In this work, a consolidated and refined WorldCup (CARWC) dataset is introduced. The dataset combines the WC14 and TSWC datasets. Additionally, all images are re-annotated with the help of publicly available custom homography refinement software developed by the principal author of this paper, which makes use of image deformation methods (Schaefer et al., 2006) to ease the annotation process and is released along with the CARWC dataset[1]. The grass band keypoints proposed in Cuevas et al. (2022) are used during annotation. These keypoints are selected since they are dense while also retaining semantic meaning. There is a total of 147 grass band keypoints across the field. In comparison, (Nie et al., 2021) uses only 91 keypoints spread uniformly across the field. The grass band keypoints are placed in semantically meaningful locations, which could aid in their detection. While uniform keypoints do not necessarily occur at the intersections of lines or other distinguishable field markings, the grass band keypoints mainly occur at the intersection of grass bands with some other field marking (which may be extended, e.g. the horizontal lines of the penalty box, with some exceptions). This also makes these keypoints easier to identify during annotation. Fig. 2(b) and Fig. 3(b) illustrate two examples of the refined annotations.

The ground truth keypoint position labels are also included along with the homography annotations. This could allow for the investigation of image distortion in future work.

---

[1] https://github.com/Paulkie99/KeypointAnnotator.

To obtain a sense of scale for the process covariance matrices of the CARWC training set, the mean matrix entries over all j of $\Sigma^{I,j}$, estimated as described in 5.1, are

$$\begin{bmatrix} 4.95 & -0.06 \\ -0.06 & 0.95 \end{bmatrix}.$$

Similarly, for Σ^H, the mean entries are

$$\begin{bmatrix} 3.35 & 0.05 & 0.00 & 1.58 & 0.27 & 0.00 & -373 & 12.77 \\ 0.05 & 0.01 & 0.00 & 0.02 & 0.00 & 0.00 & -4.66 & -0.08 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.10 & 0.00 \\ 1.58 & 0.02 & 0.00 & 0.77 & 0.13 & 0.00 & -179 & 6.45 \\ 0.27 & 0.00 & 0.00 & 0.13 & 0.03 & 0.00 & -30.11 & 0.94 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.05 & 0.00 \\ -373 & -4.66 & -0.10 & -179 & -30.11 & 0.05 & 41943 & -1458 \\ 12.77 & -0.08 & 0.00 & 6.45 & 0.94 & 0.00 & -1458 & 81.31 \end{bmatrix},$$

where covariances associated with h_{33} have been omitted because this element is not included in the Kalman filter state vector as discussed in 5.1.

5.3. Baselines

The proposed method will be evaluated primarily by comparison with two state-of-the-art keypoint-detection-based methods, namely those of Chu et al. (2022), Nie et al. (2021).

The method proposed by Nie et al. (2021) makes use of a ResNet-18-based (He et al., 2015), U-Net-like (Ronneberger et al., 2015) architecture with non-local blocks (Wang et al., 2017) and dilated convolutions. They propose the use of 91 keypoints spread uniformly across the field. In a multi-task learning approach, their method simultaneously predicts keypoints and dense features defined as the normalised distance of non-line or non-region pixels to the nearest line or region pixel in the image (referred to as line and region features, respectively). Finally, they use an online refinement scheme that considers these predicted features and the homography estimates for the current and previous frames. The re-implementation of their method in Chu et al. (2022) is used, which does not include dense feature regression or online refinement (i.e. performs keypoint detection only), and expanded with dense feature regression and online refinement, where the hyper-parameters for online refinement are set as in Nie et al. (2021). Fig. 4 shows the keypoint error distribution of the trained model on the CARWC training set — it is clear that the zero-mean Gaussian assumption is reasonable. The median matrix entries over all j of $\Sigma^{M,j}$, estimated as described in 5.1 using the method of Nie et al. (2021) on the CARWC training set, are

$$\begin{bmatrix} 20.81 & -0.01 \\ -0.01 & 14.56 \end{bmatrix}.$$

Similarly, the mean entries of the estimated covariance matrix of the homography obtained with RANSAC (used to initialise the homography elements of the Kalman filter state vector) are

$$\begin{bmatrix} 254 & 1.76 & 0.06 & 123 & 22.81 & -0.03 & -28795 & 947 \\ 1.76 & 0.25 & 0.00 & 0.54 & 0.05 & 0.00 & -163 & -1.17 \\ 0.06 & 0.00 & 0.00 & 0.03 & 0.01 & 0.00 & -6.51 & 0.21 \\ 123 & 0.54 & 0.03 & 60.23 & 11.22 & -0.02 & -14023 & 472 \\ 22.81 & 0.05 & 0.01 & 11.22 & 2.18 & 0.00 & -2605 & 87.42 \\ -0.03 & 0.00 & 0.00 & -0.02 & 0.00 & 0.00 & 3.85 & -0.12 \\ -28795 & -163 & -6.51 & -14023 & -2605 & 3.85 & 3280363 & -108856 \\ 947 & -1.17 & 0.21 & 472 & 87.42 & -0.12 & -108856 & 4186 \end{bmatrix},$$

where the elements associated with h_{33} have been omitted as in 5.2.3.

The method proposed by Chu et al. (2022), referred to as KpSFR, uses a ResNet-34-based encoder–decoder architecture incorporating skip connections. Using the same keypoint template proposed by Nie et al. (2021), dynamic filter learning is used to predict keypoints. While state-of-the-art results are reported, the method requires pre-processed results, e.g. acquired by the method of Nie et al. (2021), to obtain keypoint identity encodings during inference. Ignoring this pre-requisite,
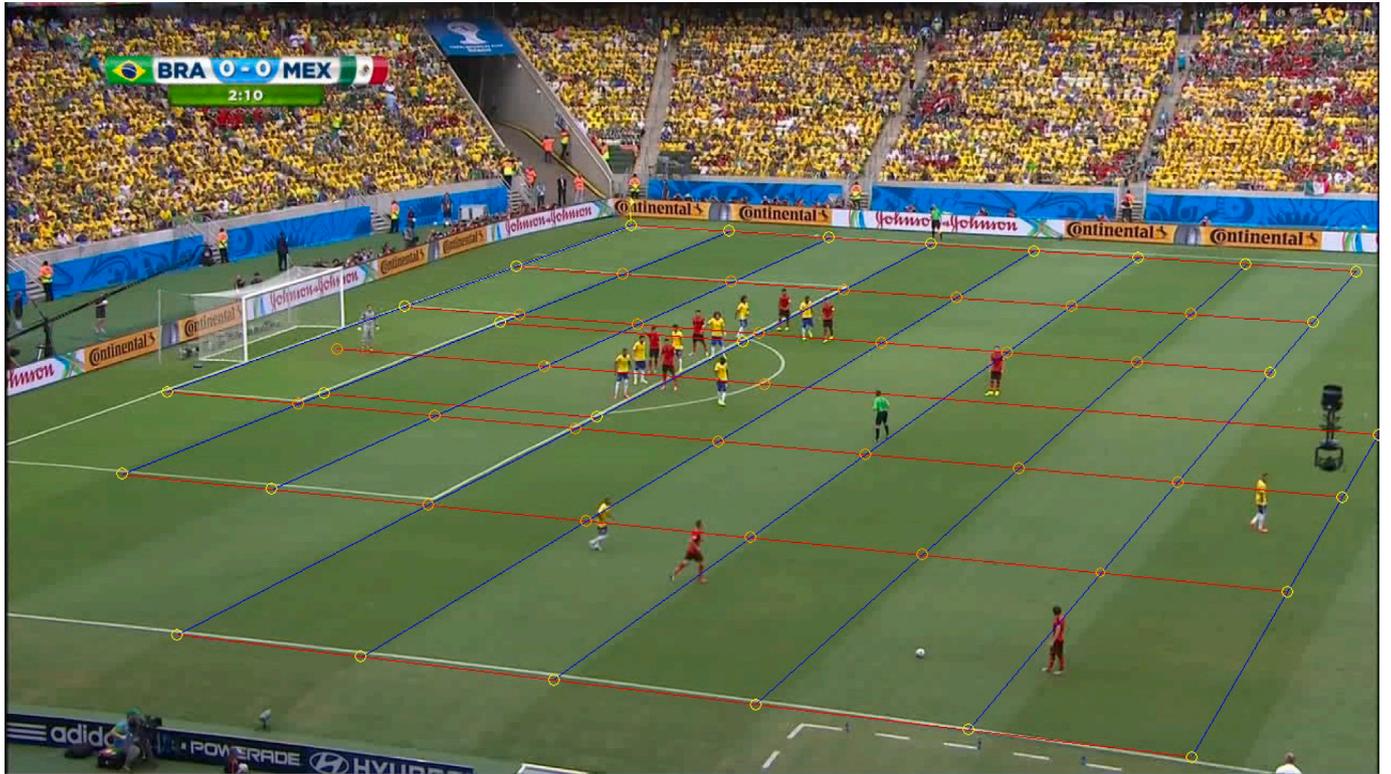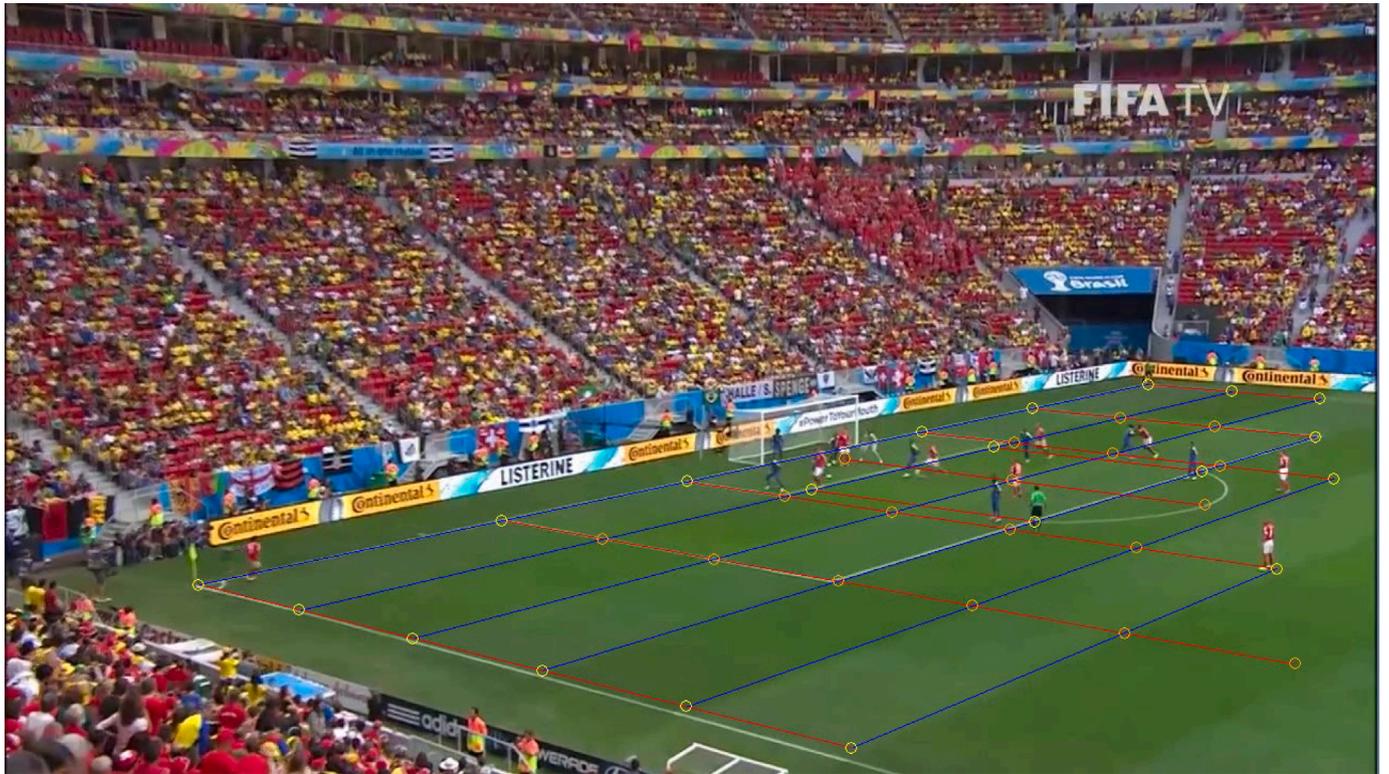
(a) WC14.
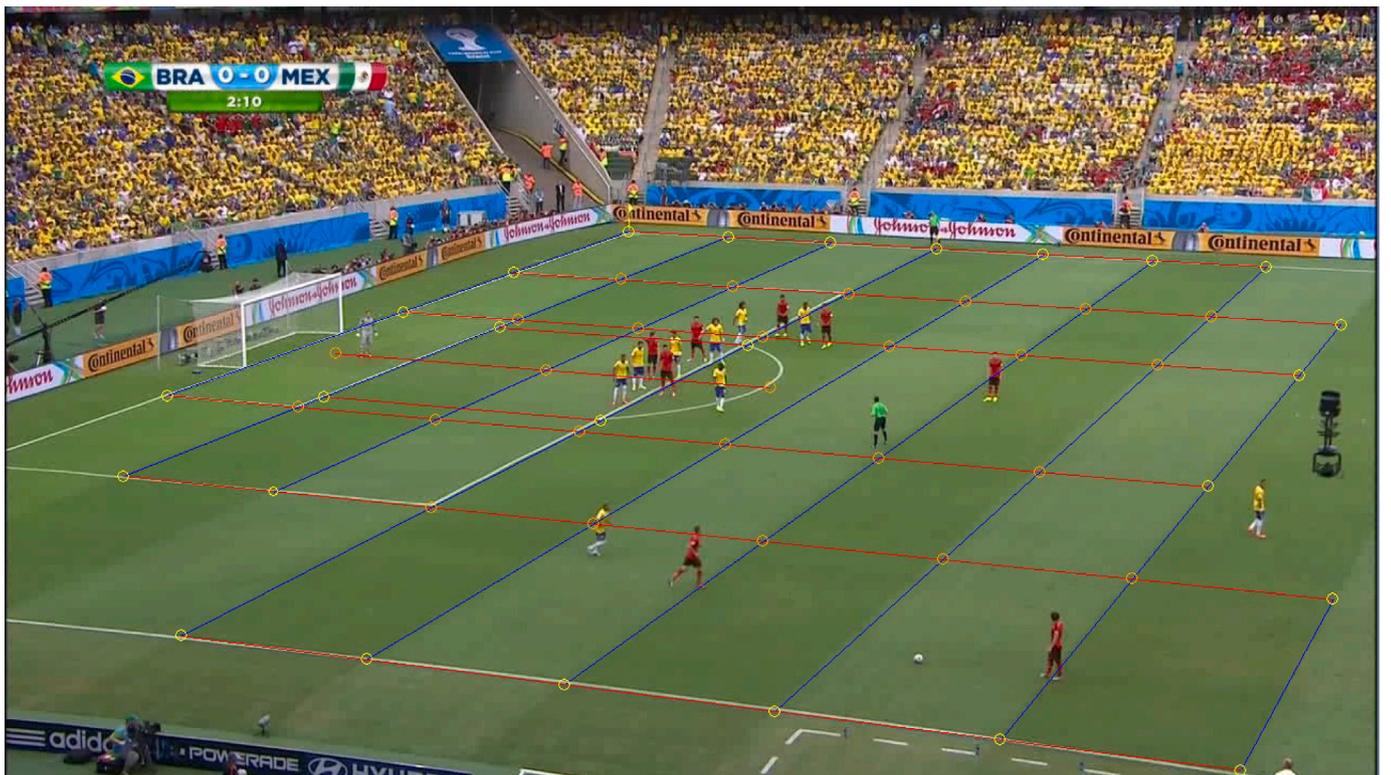
(b) CARWC.

Fig. 2. Re-projected keypoints using WC14 and custom (CARWC) annotated homographies. The re-projection error of the WC14 annotation is most noticeable when considering the alignment with the right-most grass band.

(a) TSWC.

(b) CARWC.

Fig. 3. Re-projected keypoints using TSWC and custom (CARWC) annotated homographies. The re-projection error of the TSWC annotation is most noticeable when considering the alignment with the bottom length-wise horizontal field line.
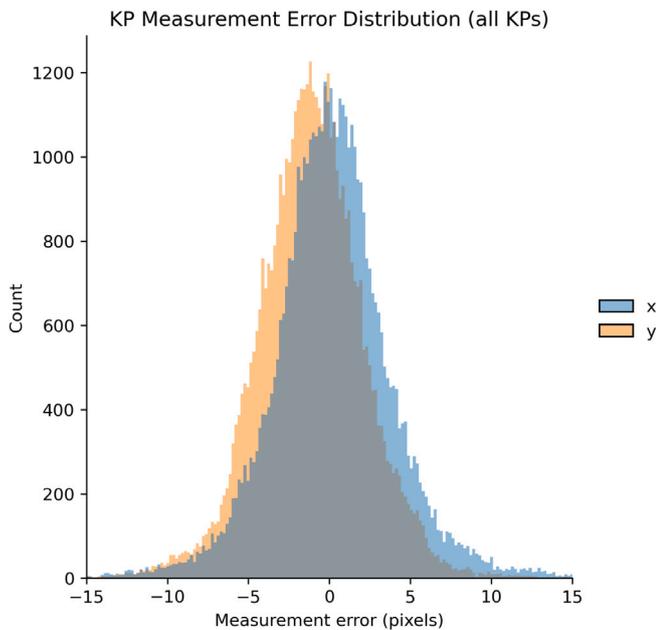
Fig. 4. Keypoint (KP) measurement error distribution in the x- and y dimensions with the baseline model of Nie et al. (2021), trained and evaluated on the CARWC training set.

the method uses approximately 73 million parameters, and inference occurs at approximately 1.5 frames per second[2]. In comparison, the re-implementation of Nie et al. (2021) requires approximately 42 million parameters and executes at approximately 50 frames per second (Nie et al., 2021).

Both of these networks are trained in a manner similar to Nie et al. (2021): with the Adam (Kingma & Ba, 2014) optimiser ($\beta_1 = 0.9, \beta_2 = 0.999$) for 300 epochs, where the initial learning rate of $1e-4$ decays to $1e-5$ after 200 epochs. Training takes place on the CARWC training set.

The performance obtained before and after augmentation with our proposed method is of particular interest. Variations of the method of Nie et al. (2021) (brought about by changes to the training programme, the number or distribution of keypoints in the field template) are augmented with our BHITK approach. While some of these variations improve over other baseline methods without BHITK, additional improvements are attained using BHITK. Thus, the Bayesian modelling approach enables performance improvements which may not be attainable through other means.

It has been shown that using stochastic gradient descent (SGD) instead of Adam leads to better generalisation (Zhou et al., 2020). Furthermore, sharpness-aware minimisation (SAM) (Foret et al., 2020) has been proposed to avoid sharp local minima, improving generalisation. The first variation of Nie et al. (2021) uses SAM and SGD instead of Adam. The learning rate for SGD is set to 0.1, with a momentum of 0.9. Adaptive SAM is used with $\rho = 2$.

The keypoint layout, i.e. the number and distribution of the keypoints in the field template, affects performance (Nie et al., 2021). Another variation, therefore, uses the grass band keypoints (Cuevas et al., 2022). To investigate the effect of uniform versus non-uniform keypoint layouts, where the total number of keypoints remains constant, yet another variation considers an increased number of uniform keypoints, such that the total matches that of the grass band keypoints (i.e. 147 keypoints with a uniform spatial distribution).

---

[2] As measured on an NVIDIA RTX 3090 GPU with an AMD Ryzen 9 5900X CPU, with the following inference script: https://github.com/ericsujw/KpSFR/blob/main/inference.py.

5.4. Introduction to evaluation metrics

Following Chu et al. (2022), Citraro et al. (2020), Nie et al. (2021), the mean and median of various evaluation metrics are reported in Section 6. These metrics are briefly explained, followed by more detailed explanations in the following subsections.

5.4.1. Homography evaluation metrics

Two types of IoU metrics evaluate the homographic projections of the video frame and the field template, respectively. Additionally, the projection error of randomly sampled points in the video frame projected onto the field template and the re-projection error of the field template keypoints into the video frame are reported. Re-projection refers to transforming a coordinate in the field template to a point in the video frame using the ground truth or estimated (predicted) homography. Projection refers to the inverse of this transformation.

5.4.2. Keypoint measurement metrics

The normalised root-mean-square errors (NRMSE) for keypoint coordinates in the x and y image dimensions are reported to evaluate image keypoint position estimates. Furthermore, precision and recall are reported for a given distance threshold, while the mean average precision (mAP) is used to evaluate keypoint detections over a range of distance thresholds.

5.5. Evaluation metrics

5.5.1. Intersection over union

The first type of IoU considered, IoU_{entire}, is obtained by re-projecting the field template mask – i.e. the rectangle that represents the soccer field – using the ground truth homography. This re-projection is then projected using the predicted homography. The IoU_{entire} is equal to the area of intersection of this projected polygon and the field template polygon, divided by the area of their union. I.e. if the estimated homography performs the same mapping as that of the ground truth over the entire field, $IoU_{entire} = 1$. This definition of IoU_{entire} is consistent with Citraro et al. (2020), Nie et al. (2021). However, the IoU_{entire} is calculated incorrectly in Chu et al. (2022). Instead of using the field template mask, Chu et al. (2022) projects the binary mask representing the image area onto the field template using the ground truth homography. The resulting polygon is then re-projected to image coordinates using the predicted homography. The IoU_{entire} is obtained as the area of intersection of the re-projected polygon and the binary mask representing the image area, divided by the area of their union[3].

The second type of IoU, IoU_{part}, considers only the part of the field which is visible in the video frame. The ground truth homography projects the binary mask representing the image area. The predicted homography also performs this projection. Each of these projections results in a polygon in the field template. The IoU_{part} is equal to the area of intersection of these polygons divided by their area of union.

A point of concern in Citraro et al. (2020) is that IoU_{part} does not take the mapping accuracy outside of the visible field into account. Nevertheless, it was pointed out in Nie et al. (2021) that the ground truth homography is determined only from the visible part of the field. Therefore, IoU_{entire} is not necessarily reliable since the ground truth mapping is not guaranteed to be accurate outside of the visible field. These concerns are valid, and annotations are never perfect. However, with the re-annotation of the WC14 and TSWC datasets, resulting in the CARWC dataset, it is believed that both of these metrics are more reliable.

---

[3] https://github.com/ericsujw/KpSFR/blob/main/metrics.py.

Table 2

A comparison of the homography evaluation performance of the baseline detection methods with and without BHITK on the time-series part of the CARWC dataset. Models augmented with BHITK are marked with a checkmark in the BHITK column.

Method	BHITK	IoU_{entire} (%) ↑		IoU_{part}(%) ↑		Proj. (meter) ↓		Re-Proj. (%)↓	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
Nie et al. (2021) with online refinement	X	86.76	89.67	98.18	98.43	0.38	0.35	0.89	0.78
Nie et al. (2021)	X	86.79	89.67	98.19	98.43	0.37	0.35	0.88	0.78
	✓	**90.48**	**92.14**	**98.63**	**98.88**	**0.34**	**0.33**	**0.77**	**0.73**
Nie et al. (2021) + SAM + SGD	X	85.77	87.93	98.37	98.64	0.32	0.28	0.77	0.71
	✓	**92.29**	**93.68**	**98.87**	**99.00**	**0.25**	**0.22**	**0.59**	**0.57**
Nie et al. (2021) + SAM + SGD + more KPs	X	85.08	87.31	98.48	98.67	0.30	0.28	0.76	0.68
	✓	**91.33**	**92.99**	**98.89**	**99.06**	**0.26**	**0.23**	**0.61**	**0.52**
Nie et al. (2021) + SAM + SGD + grass band KPs	X	84.65	88.26	98.52	98.66	0.30	0.28	0.70	0.66
	✓	**91.69**	**93.68**	**98.94**	**99.08**	**0.23**	**0.23**	**0.55**	**0.53**
KpSFR (Chu et al., 2022)	X	89.52	91.40	98.36	98.68	0.42	0.37	0.83	0.71

5.5.2. Projection error

The projection error is calculated similarly to Chu et al. (2022), Nie et al. (2021). First, 2500 image points are randomly sampled from a uniform distribution defined over the part of the image where the soccer field is visible. This image portion is determined by the re-projection of the field template mask using the ground truth homography. The projection error in meters is then calculated as the average pair-wise distance between the projections of these points using the predicted homography and their projections using the ground truth homography. However, whereas Chu et al. (2022), Nie et al. (2021) assumed field dimensions of 100×60 m, dimensions of 105×68 m are used instead (unless otherwise specified) — which more accurately reflect FIFA regulations (FIFA, 2022).

5.5.3. Re-projection error

Re-projection error is the average pair-wise distance between the field template keypoints re-projected in the image using the ground truth homography and those re-projected using the predicted homography, normalised by the image height (Chu et al., 2022; Nie et al., 2021).

5.5.4. Normalised root-mean-square error

To investigate the effect of keypoint filtering on keypoint position estimates, the NRMSE is used:

$$\text{NRMSE} = \frac{1}{Z\sqrt{L}} \sqrt{\sum_{l=1}^{L} \left(x_l^{I,j} - \hat{x}_l^{I,j} \right)^2},$$

where L is the total number of measured keypoints which correspond to the ground truth, $x_l^{I,j}$ and $\hat{x}_l^{I,j}$ are the corresponding ground truth and estimated keypoint positions, respectively, in either the x- or y-dimension. Finally, Z is the image width or height corresponding to the computation of the NRMSE in the x- or y-dimension.

5.5.5. Precision, recall and mean-average precision

Precision is the ratio of true positive detections to the number of predicted detections, while recall is the ratio of true positive detections to the number of ground truth detections. Following Chu et al. (2022), a keypoint is considered a true positive if it is within a distance of 5 pixels to the ground truth position in the predicted image space (320×180), which is equivalent to 20 pixels in the actual image space (1280×720) or $\sim 2.78\%$ of the image height. Additionally, the average precision (AP) is evaluated at 5, 10, 15 and 20-pixel thresholds (in the actual image space):

$$\text{AP} = \sum_{n} \left(R_n - R_{n-1} \right) P_n,$$

where R_n and P_n are the recall and precision at the nth threshold. The mAP is then obtained as the mean AP.

6. Results and discussion

The following evaluation metrics are obtained by performing inference on the CARWC test set unless specified otherwise.

6.1. Baseline results

The results of the baseline methods are presented in Table 2 and Table 3, where they are marked with a cross in the BHITK column. Table 2 shows the homography evaluation metrics, and Table 3 shows the keypoint detection and measurement metrics.

The online refinement algorithm using two loss functions proposed by Nie et al. (2021) did not improve the results. This may be because the self-verification step, which must fail for the online refinement to occur, is consistently successful. In other words, according to the criteria of the online refinement algorithm, the estimated homography is sufficient for most of the CARWC test set. This is consistent with the results presented in Nie et al. (2021), where this refinement algorithm had an insignificant impact on the WC14 dataset.

The use of SAM and SGD improves every metric when compared to the network trained with Adam (except IoU_{entire}), thus confirming their positive effect on the generalisation of keypoint detection. With SAM and SGD, recall increases from 90.59% to 95.10%, and mAP increases from 61.70% to 68.58%. The NRMSE, projection error and re-projection error are also lowered significantly. Despite the improvements to the keypoint detection metrics and IoU_{part}, the performance of IoU_{entire} is degraded compared to the network trained with Adam. This is because RANSAC (Fischler & Bolles, 1987) only considers keypoints visible in the current frame, with no mechanism to maintain consistency in the homography between subsequent frames. Thus, it is possible for the out-of-frame mapping to be inconsistent while the within-frame mapping improves. Another explanation could be that the detection model prefers specific keypoints over others. This preference could be due to similarities between the preferred and training data keypoints, while the undetected or less preferred keypoints may be dissimilar. Thus, it is possible that keypoints which may have been essential to obtain an accurate IoU_{entire} are missed. However, this explanation is less likely since the keypoint detection metrics (specifically recall) are relatively high.

Adding more uniform keypoints to the field template slightly improves the IoU_{part}, projection and re-projection metrics but degrades the keypoint detection metrics and IoU_{entire}. The increased dimensionality of the output may explain the degradation in detection metrics. Nevertheless, this is not sufficient to negate the positive effect of having an increased number of detected keypoints on IoU_{part}.

The use of grass band keypoints results in the highest precision, lowest re-projection error, and highest mean IoU_{part} obtained amongst all the baseline methods. However, it may be concluded from the more

Table 3

A comparison of the keypoint detection and measurement performance of the baseline detection methods with and without BHITK on the time-series part of the CARWC dataset. Models augmented with BHITK are marked with a checkmark in the BHITK column.

Method	BHITK	NRMSE (%)↓		P(%)↑	R(&)↑	mAP(%)↑
		y	x			
Nie et al. (2021) with online refinement	X	0.65	0.71	94.98	90.59	61.70
Nie et al. (2021)	X	0.65	0.71	94.98	90.59	61.70
	✓	**0.63**	**0.68**	**95.49**	**91.08**	**62.26**
Nie et al. (2021) + SAM + SGD	X	0.53	0.57	96.14	95.10	68.58
	✓	**0.50**	**0.55**	**96.42**	**95.36**	**69.97**
Nie et al. (2021) + SAM + SGD + more KPs	X	0.59	0.58	95.95	94.96	66.77
	✓	**0.55**	**0.56**	**96.23**	**95.24**	**67.87**
Nie et al. (2021) + SAM + SGD + grass band KPs	X	0.63	0.58	96.44	94.86	65.39
	✓	**0.58**	**0.55**	**96.95**	**95.36**	**66.37**
KpSFR (Chu et al., 2022)	X	0.67	0.66	95.13	93.28	66.58

Table 4

The percentage improvement of the best performance metrics obtained with BHITK, relative to those obtained without BHITK.

IoU$_{entire}$		IoU$_{part}$		Proj. (meter)		Re-Proj.		NRMSE		P	R	mAP
Mean	Median	Mean	Median	Mean	Median	Mean	Median	y	x			
3.09%	**2.49**%	**0.43**%	**0.41**%	**23.33**%	**21.43**%	**21.43**%	**21.21**%	**5.66**%	**3.51**%	**0.53**%	**0.27**%	**2.03**%

considerable difference between the mean and median IoU$_{entire}$ (3.61%) that there are more outlier homography estimates when the entire field is taken into consideration. Although detections are more precise, possibly due to keypoints being placed in more semantically meaningful locations, recall and mAP are lower than for uniform keypoints.

Interestingly, KpSFR (Chu et al., 2022) obtains the worst projection error and fairs quite poorly when considering the re-projection and keypoint detection metrics. Nevertheless, it achieves the highest IoU$_{entire}$ metrics by a comfortable margin. It slightly improves upon the best median IoU$_{part}$ achieved by the variations of Nie et al. (2021).

6.2. Results with the proposed method

The results of augmenting the baseline methods with the proposed approach are in Table 2 and Table 3, marked with a checkmark in the BHITK column. For each distinct variation of Nie et al. (2021), all evaluation metrics improve using BHITK. The IoU$_{entire}$ benefits significantly from homography filtering: the large difference between the mean and median of IoU$_{entire}$ for the variation which uses grass band keypoints has been reduced from 3.61% to 1.99% (i.e. there are fewer outlier homography estimates compared to the baseline). Furthermore, the mean IoU$_{entire}$ increased by 7.04% for this variation. Relative to the unaugmented mean IoU$_{entire}$, this is an improvement of 8.32%. Significant improvements of the IoU$_{entire}$ metrics are also seen for the other baseline methods. While IoU$_{part}$ also improves with BHITK, these improvements are less striking since the IoU$_{part}$ obtained without BHITK is already relatively high. The best of the unaugmented mean and median projection errors were reduced by 7 cm and 6 cm, respectively, with BHITK. Similarly, the best re-projection errors were improved by 0.15% and 0.14% (when compared to the best of these metrics obtained with BHITK over all the experiments). These improvements may seem marginal, but the percentage improvement relative to the results obtained with the unaugmented methods is significant, as shown in Table 4.

With BHITK, the best mAP increased from 68.58% to 69.97%, the best recall from 95.10% to 95.36% and the best precision from 96.44% to 96.95%. There are also minor improvements to the best of the NRMSE metrics. This shows that the keypoint positions are effectively refined, contributing to the homography filter's effectiveness. However, the low mAP metrics relative to the generally high precision and recall suggest that keypoint identification still suffers at lower thresholds.

Augmenting (Nie et al., 2021), with no variations, with BHITK improves nearly all homography evaluation metrics over those achieved by KpSFR (Chu et al., 2022), except for the median re-projection error where the difference is only 0.02% (a relative degradation of 2.82%). This is despite having a much lower mAP, recall and slightly higher NRMSE in the x dimension. The improvement is especially significant considering the increased parameter count (73 million) and inference time (1.5 frames per second) of KpSFR compared to Nie et al. (2021) (42 million and 50 frames per second, respectively). Thus, BHITK enables a less sophisticated and less computationally expensive method to outperform the state-of-the-art in most homography evaluation metrics. Furthermore, these improvements are obtained without necessarily altering the existing method. Finally, BHITK may enable performance that is not attainable by considering such alterations, as is shown by the fact that each distinct variation improved significantly using BHITK.

6.3. Results on TSWC

The proposed method is applied to a variation of Nie et al. (2021) without dense feature regression, as implemented in Chu et al. (2022), to compare results on the TSWC dataset. Table 5 shows that the improvement afforded by the proposed method is consistent across the different datasets. The simplified (Nie et al., 2021) network with BHITK outperforms or achieves similar homography metrics to more computationally expensive methods, such as KpSFR (Chu et al., 2022) and the method of Chen and Little (2018). These improvements are achieved despite a much lower recall than that of KpSFR — similar to the improvement of Nie et al. (2021) over KpSFR despite lower recall in 6.2. Most keypoint detection metrics also do not show a significant improvement. Thus, it may be concluded that the homography filter plays the most significant role in the proposed method.

6.4. Qualitative evaluations

Qualitative comparisons are shown in Figs. 5 and 6. In each figure, the homography estimate predicted for the same image is used to re-project the field template keypoints into the image and project the image onto the field template. In each case, the first sub-figure represents the results obtained without using BHITK, while the second shows the results using BHITK. Specifically, Fig. 5(a) shows the results obtained with the homography estimate obtained from the method of Nie et al. (2021), while Fig. 5(b) shows the results obtained when the same method is augmented with BHITK. Similarly, Fig. 6(a) shows the results obtained with the variation of Nie et al. (2021) trained with
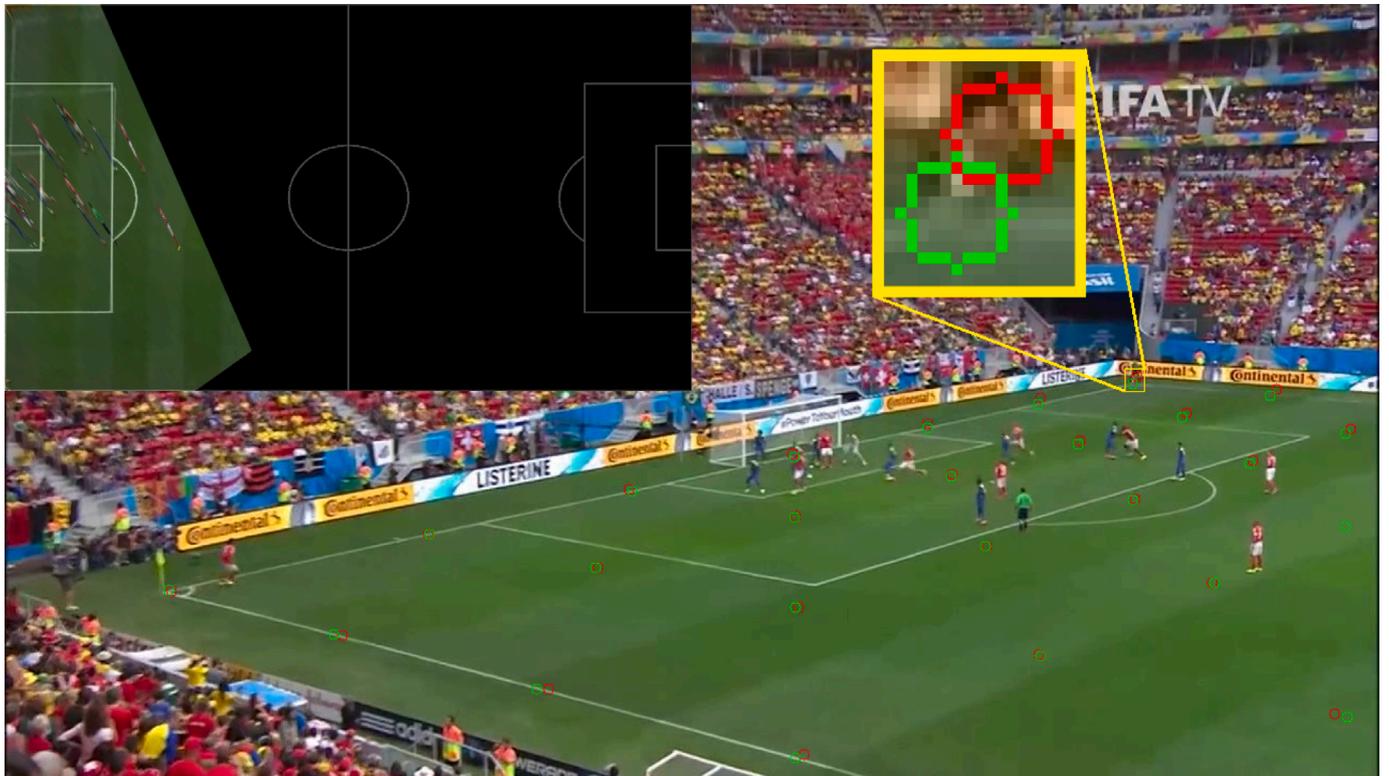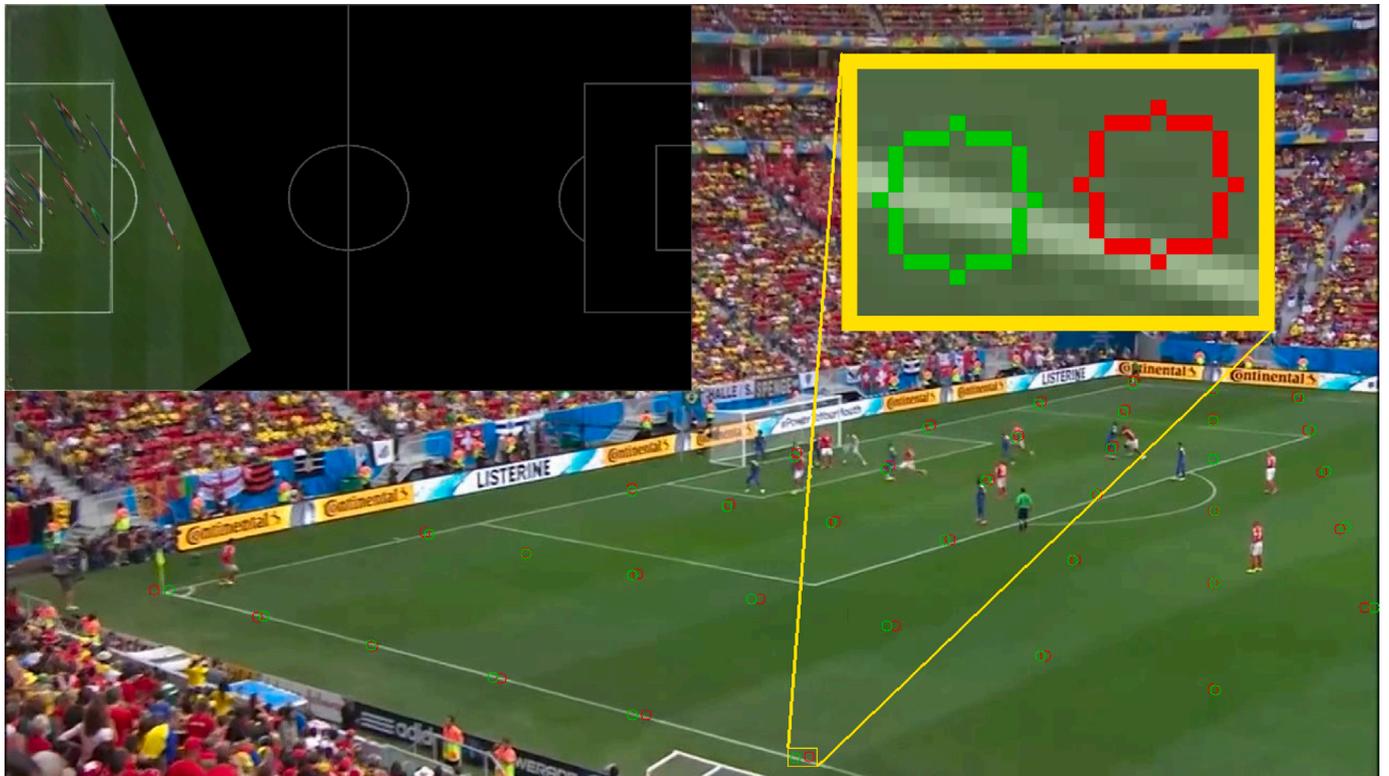
(a) Projection and re-projection using the homography estimate obtained with (Nie et al., 2021).
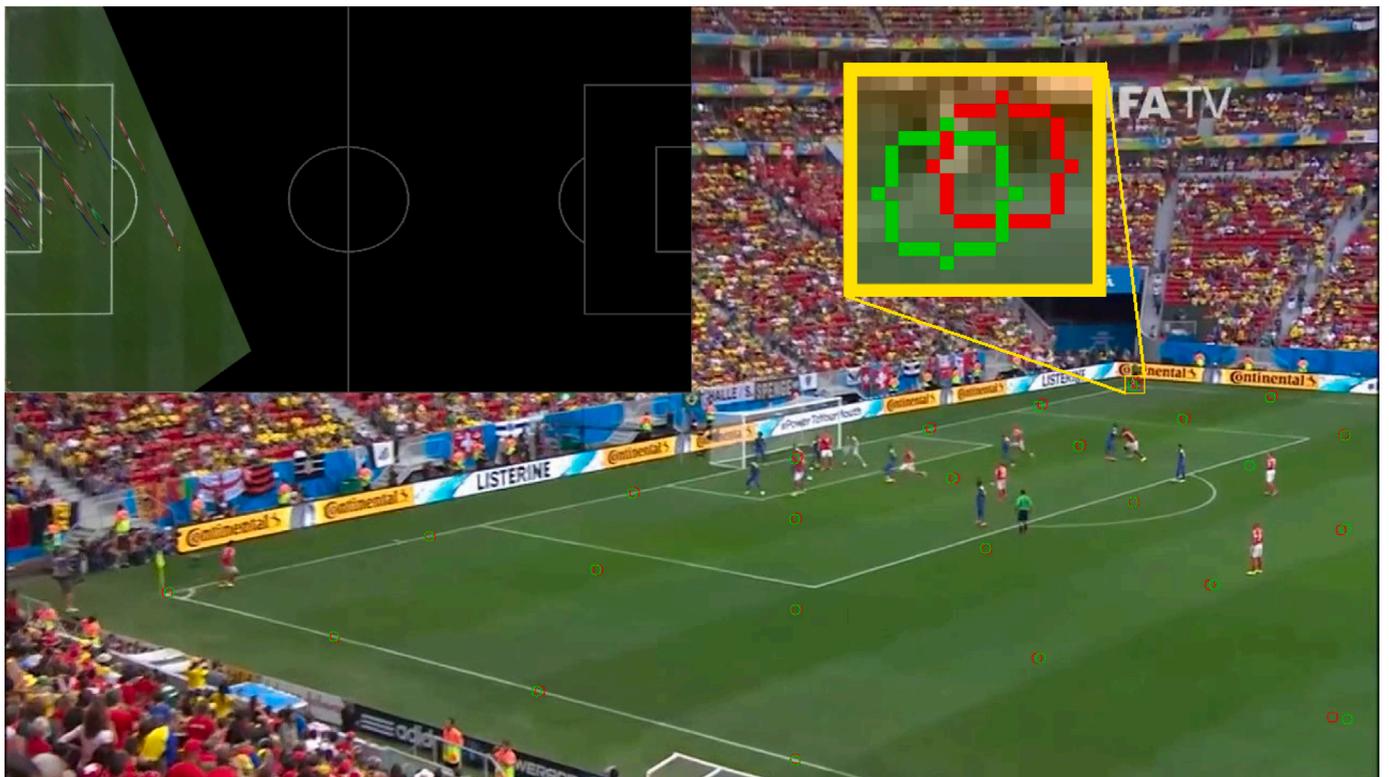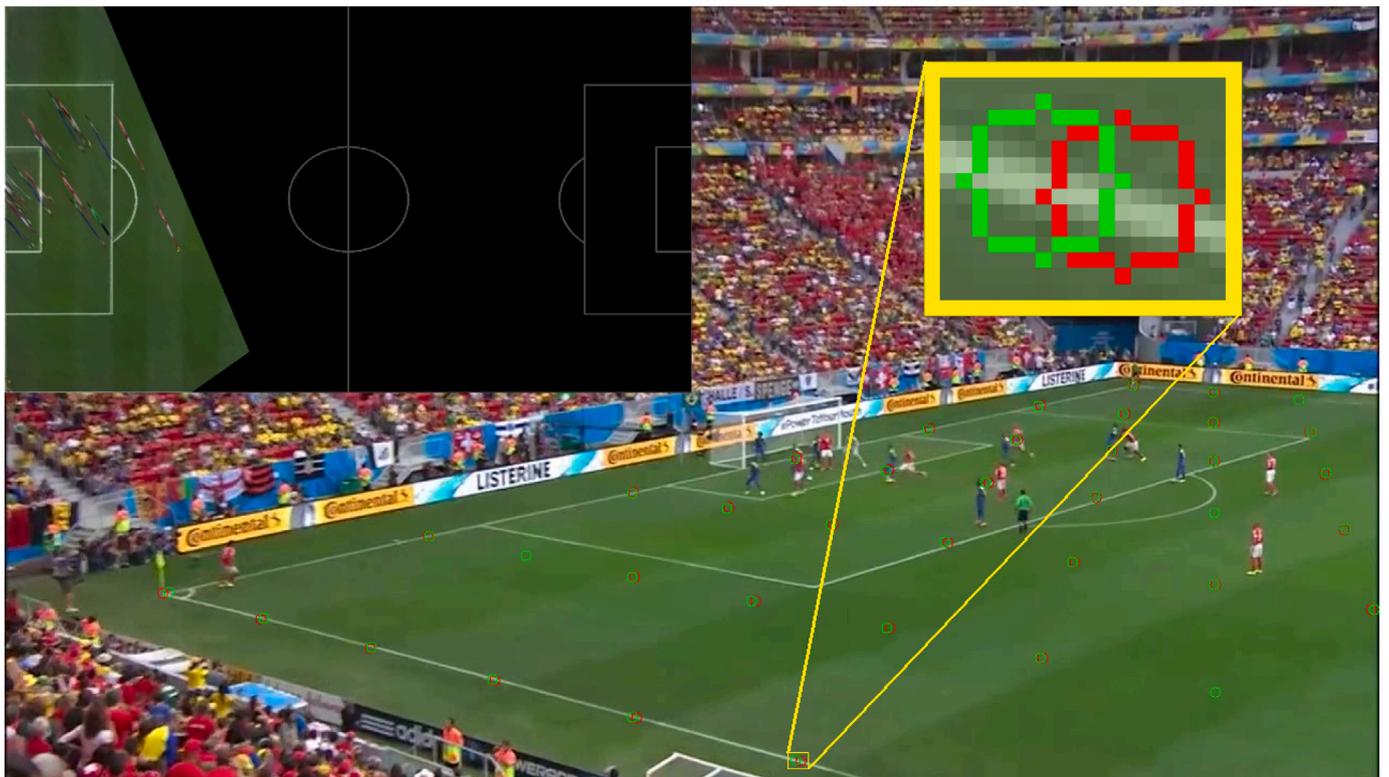

(b) Projection and re-projection using the homography estimate obtained with (Nie et al., 2021) + BHITK.

Fig. 5. This figure depicts both the re-projected keypoints within the image and the image projected onto the field template. The projection and re-projection in each sub-figure utilise distinct homography estimates: one derived from Nie et al. (2021)'s method and the other obtained by augmenting this same network with the proposed method (BHITK). The sub-figures represent the same frame from the same test video. The red circles represent the keypoints re-projected using the predicted homography, and the green circles represent the keypoints re-projected using the ground truth homography.

(a) Projection and re-projection using the homography estimate obtained with (Nie et al., 2021) + SAM + SGD + more KPs.

(b) Projection and re-projection using the homography estimate obtained with (Nie et al., 2021) + SAM + SGD + more KPs + BHITK.

Fig. 6. This figure depicts both the re-projected keypoints within the image and the image projected onto the field template. The projection and re-projection in each sub-figure utilise distinct homography estimates: one derived from a variant of Nie et al. (2021)'s method, which is trained with SAM, SGD and 147 keypoints, and the other obtained by augmenting this same network with the proposed method (BHITK). The sub-figures represent the same frame from the same test video. The red circles represent the keypoints re-projected using the predicted homography, and the green circles represent the keypoints re-projected using the ground truth homography.

Table 5

A comparison of BHITK with past methods on the TSWC dataset.

Method	IoU$_{entire}$ (%) ↑		IoU$_{part}$(%) ↑		Proj. (meter) ↓		Re-Proj. (%)↓		NRMSE (%) ↓		P(%)↑	R(%)↑	mAP(%)↑
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	y	x			
Chen and Little (2018) as reported in Chu et al. (2022)	90.7[a]	94.1[a]	96.8	97.4	0.54[b]	0.38[b]	1.6	1.3	–	–	–	–	–
Nie et al. (2021) as in Chu et al. (2022)	92.5[a]	94.2[a]	97.4	97.9	0.43[b]	0.37[b]	1.1	1.0	0.66	0.78	94.96	83.23	56.87
KpSFR (Chu et al., 2022)	94.8[a]	95.4[a]	98.1	98.2	0.36[b]	0.33[b]	0.9	0.8	–	–	–	87	–
Nie et al. (2021) as in Chu et al. (2022) + **BHITK**	**94.8**[a]	**95.8**[a]	**97.9**	**98.3**	**0.36**[b]	**0.32**[b]	**0.9**	**0.8**	**0.62**	**0.76**	**95.06**	**83.32**	**56.92**

[a] The IoU$_{entire}$ metric is calculated as in Chu et al. (2022).

[b] The projection error is calculated using field template dimensions of 100 × 60 m, as in Chu et al. (2022).

SAM, SGD and 147 keypoints, while Fig. 6(b) shows the results when this variation is augmented with BHITK.

In Fig. 5 and Fig. 6, the re-projected keypoints are closer to their ground truth positions when BHITK is employed. Furthermore, the projected field lines align more closely with those of the field template when using BHITK.

7. Conclusion

Exploiting the temporally consistent nature of homographic projections in a Bayesian framework is shown to be beneficial. The proposed approach effectively enforces temporal consistency between subsequent homography estimates through an affine transformation. As long as the underlying keypoint detection method satisfies the standard Kalman filter assumptions (i.e. approximately zero-mean Gaussian-distributed measurement noise), homography filtering from tracked keypoints is shown to be effective. When augmented with the proposed method, the overall weakest-performing baseline method outperforms the state-of-the-art, which is much more computationally expensive, in all but one of the homography evaluation metrics (median re-projection error, where the difference is only 0.02%). Furthermore, all baseline evaluation metrics improve when the baseline methods are augmented with BHITK. Thus, the method will likely improve the performance of several existing keypoint detection methods. Finally, the annotations of the WorldCup and TS-WorldCup datasets are refined and released along with a custom homography annotation tool as the CARWC dataset.

CRediT authorship contribution statement

Paul Johannes Claasen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Johan Pieter de Villiers:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Paul Claasen reports financial support was provided by MultiChoice. Paul Claasen reports a relationship with MultiChoice that includes: consulting or advisory, employment, and funding grants.

Data availability

While the datasets used are publicly available, the code cannot be shared at this time.

Acknowledgements

References

Aharon, N., Orfaig, R., & Bobrovsky, B.-Z. (2022). BoT-SORT: Robust associations multi-pedestrian tracking. ArXiv, arXiv:2206.14651.

Bergmann, P., Meinhardt, T., & Leal-Taixé, L. (2019). Tracking without bells and whistles. In *2019 IEEE/CVF international conference on computer vision* (pp. 941–951).

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Brown, M., & Lowe, D. G. (2007). Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, *74*(1), 59–73. http://dx.doi.org/10.1007/s11263-006-0002-3.

Chen, J., & Little, J. (2018). Sports camera calibration via synthetic data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 2497–2504).

Chu, Y. J., Su, J. W., Hsiao, K. W., Lien, C. Y., Fan, S. H., Hu, M. C., Lee, R. R., Yao, C. Y., & Chu, H. K. (2022). Sports field registration via keypoints-aware label condition. In *2022 IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 3522–3529). http://dx.doi.org/10.1109/CVPRW56347.2022.00396.

Citraro, L., Márquez-Neila, P., Savarè, S., Jayaram, V., Dubout, C., Renaut, F., Hasfura, A., Ben Shitrit, H., & Fua, P. (2020). Real-time camera pose estimation for sports fields. *Machine Vision and Applications*, *31*(3), 16. http://dx.doi.org/10.1007/s00138-020-01064-7.

Cuevas, C., Berjón, D., & García, N. (2022). Grass band detection in Soccer images for improved image registration. *Signal Processing: Image Communication*, *109*, Article 116837. http://dx.doi.org/10.1016/j.image.2022.116837, URL https://www.sciencedirect.com/science/article/pii/S0923596522001266.

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2016). Deep image homography estimation. ArXiv arXiv:1606.03798.

Du, Y., Wan, J.-J., Zhao, Y., Zhang, B., Tong, Z., & Dong, J. (2021). GIAOTracker: A comprehensive framework for MCMOT with global information and optimizing strategies in VisDrone 2021. In *2021 IEEE/CVF international conference on computer vision workshops* (pp. 2809–2819).

Evangelidis, G. D., & Psarakis, E. Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(10), 1858–1865. http://dx.doi.org/10.1109/TPAMI.2008.113.

FIFA (2022). Stadium guidelines 5.3 Pitch dimensions and surrounding areas. URL https://publications.fifa.com/en/football-stadiums-guidelines/technical-guideline/stadium-guidelines/pitch-dimensions-and-surrounding-areas/.

Fischler, M. A., & Bolles, R. C. (1987). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In M. A. Fischler, & O. Firschein (Eds.), *Readings in computer vision* (pp. 726–740). San Francisco (CA): Morgan Kaufmann, http://dx.doi.org/10.1016/B978-0-08-051581-6.50070-2, URL https://www.sciencedirect.com/science/article/pii/B9780080515816500702.

Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2020). Sharpness-aware minimization for efficiently improving generalization. ArXiv, arXiv:2010.01412.

Han, S., Huang, P., Wang, H., Yu, E., Liu, D., Pan, X., & Zhao, J. (2020). MAT: Motion-aware multi-object tracking. ArXiv, arXiv:2009.04794.

Hartley, R., & Zisserman, A. (2004). *Multiple view geometry in computer vision* (2nd ed.). Cambridge: Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511811685, URL https://www.cambridge.org/core/books/multiple-view-geometry-in-computer-vision/0B6F289C78B2B23F596CAA76D3D43F7A.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Homayounfar, N., Fidler, S., & Urtasun, R. (2017). Sports field localization via deep structured models. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 4012–4020). http://dx.doi.org/10.1109/CVPR.2017.427.

Isard, M., & Blake, A. (1998). CONDENSATION—Conditional density propagation for visual tracking. *International Journal of Computer Vision*, *29*(1), 5–28. http://dx.doi.org/10.1023/A:1008078328650.

Jiang, W., Higuera, J. C. G., Angles, B., Sun, W., Javan, M., & Yi, K. M. (2019). Optimizing through learned errors for accurate sports field registration. In *2020 IEEE Winter Conference on Applications of Computer Vision* (pp. 201–210).

Khurana, T., Dave, A., & Ramanan, D. (2020). Detecting invisible people. In *2021 IEEE/CVF international conference on computer vision* (pp. 3154–3164).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. CoRR abs/1412.6980, arXiv:1412.6980.

Le, H., Liu, F., Zhang, S., & Agarwala, A. (2020). Deep homography estimation for dynamic scenes. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 7649–7658).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

Lucas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision (IJCAI). *[no source information available]: vol. 81.*

Mur-Artal, R., Montiel, J. M. M., & Tardós, J. D. (2015). ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, *31*(5), 1147–1163. http://dx.doi.org/10.1109/TRO.2015.2463671.

Nguyen, T., Chen, S. W., Shivakumar, S. S., Taylor, C. J., & Kumar, V. (2018). Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, *3*(3), 2346–2353. http://dx.doi.org/10.1109/LRA.2018.2809549.

Nie, X., Chen, S., & Hamid, R. (2021). A robust and efficient framework for sports-field registration. In *2021 IEEE winter conference on applications of computer vision* (pp. 1935–1943). http://dx.doi.org/10.1109/WACV48630.2021.00198.

Nishida, K., Fujiki, J., Tsuchiya, C., Tanaka, S., & Kurita, T. (2011). Road plane detection using differential homography estimated by pair feature matching of local regions. In *Signal processing, pattern recognition, and applications / 722: computer graphics and imaging.* http://dx.doi.org/10.2316/P.2011.721-073.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2015* (pp. 234–241). Cham: Springer International Publishing.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *2011 international conference on computer vision* (pp. 2564–2571). http://dx.doi.org/10.1109/ICCV.2011.6126544.

Schaefer, S., McPhail, T., & Warren, J. (2006). Image deformation using moving least squares. *ACM Transactions on Graphics*, *25*, 533–540. http://dx.doi.org/10.1145/1179352.1141920.

Sha, L., Hobbs, J., Felsen, P., Wei, X., Lucey, P., & Ganguly, S. (2020). End-to-end camera calibration for broadcast videos. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 13624–13633). http://dx.doi.org/10.1109/CVPR42600.2020.01364.

Shi, F., Marchwica, P., Higuera, J. C. G., Jamieson, M., Javan, M., & Siva, P. (2022). Self-supervised shape alignment for sports field registration. In *2022 IEEE/CVF winter conference on applications of computer vision* (pp. 3768–3777). http://dx.doi.org/10.1109/WACV51458.2022.00382.

Simon, G., Fitzgibbon, A. W., & Zisserman, A. (2000). Markerless tracking using planar structures in the scene. In *Proceedings IEEE and ACM international symposium on augmented reality* (pp. 120–128). http://dx.doi.org/10.1109/ISAR.2000.880935.

Theiner, J., & Ewerth, R. (2022). TVCalib: Camera calibration for sports field registration in Soccer. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1166–1175).

Vlahović, N., Ilić, N., & Stanković, M. (2018). Deep learning in video stabilization homography estimation. In *2018 14th symposium on neural networks and applications* (pp. 1–5). http://dx.doi.org/10.1109/NEUREL.2018.8587021.

Wan, E. A., & Merwe, R. V. D. (2000). The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 adaptive systems for signal processing, communications, and control symposium (cat. no.00EX373)* (pp. 153–158). http://dx.doi.org/10.1109/ASSPCC.2000.882463.

Wang, X., Girshick, R. B., Gupta, A. K., & He, K. (2017). Non-local neural networks. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 7794–7803).

Wu, F., Wei, H., & Wang, X. (2017). Correction of image radial distortion based on division model. *Optical Engineering, 56.*

Zhang, Z. (1999). Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the seventh IEEE international conference on computer vision: vol. 1*, (pp. 666–673). http://dx.doi.org/10.1109/ICCV.1999.791289.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., & Weinan, E. (2020). Towards theoretically understanding why SGD generalizes better than ADAM in deep learning. ArXiv, arXiv:2010.05627.