FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Homography Calculation for Football Matches Using Moving Cameras

Carlos Manuel Ferreira Gomes



Master in Informatics and Computing Engineering

Supervisor: Prof. Pedro Carvalho

July 24, 2024

© Carlos Manuel Ferreira Gomes, 2024

Homography Calculation for Football Matches Using Moving Cameras

Carlos Manuel Ferreira Gomes

Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

President: Prof. Luís Filipe Teixeira
Referee: Prof. Paula Viana
Referee: Prof. Pedro Carvalho

July 24, 2024

Resumo

O futebol é um desporto altamente competitivo, onde a mais pequena diferença a nível de desempenho pode ter um impacto significativo no resultado de um jogo. Esta característica, aliada à sua popularidade mundial, atrai diversos investimentos para alcançar sucesso a nível desportivo. Esta realidade impõe às equipas a necessidade de aplicarem de forma eficiente este financiamento para aprimorar resultados. Perante os desafios, económicos e desportivos, o uso de tecnologia surge como uma solução para melhorar o desempenho e ganhar vantagem sobre rivais.

Atualmente, o processamento automático de vídeo é das principais tecnologias utilizadas para otimizar estratégias de jogo. Porém, a maioria das soluções existentes requerem *hardware* caro e complexo, como múltiplas câmaras fixas, sensores ou dispositivos GPS. Este facto proporciona um desafio significativo para muitos clubes, especialmente os que têm maior limitação a nível de recursos.

Relativamente às adversidades que surgem das soluções anteriores, a utilização de imagens e vídeos em cenários de baixas restrições, especialmente com uso de uma câmara móvel, permite que um público mais alargado beneficie de tecnologia de ponta. No entanto, uma câmara móvel introduz desafios acrescidos devido à variação de imagem causada pelo movimento, o que requer fiabilidade e adaptabilidade dos algoritmos para a análise de vídeo.

Esta investigação propõe uma abordagem para o cálculo de homografia em jogos de futebol em cenários de câmara móvel, abordando as limitações dos métodos atuais em ambientes dinâmicos e complexos. As técnicas existentes, que dependem fortemente de pontos de interesse visíveis e são aplicadas *frame* a *frame*, demonstram falhas no cálculo de homografia em zonas do campo com poucos pontos de interesse. Para superar estes desafios, propomos dois métodos: criar uma super imagem panorâmica do vídeo utilizando o movimento da câmara e integrar os dados de movimento diretamente nas equações de homografia. Foi utilizado o fluxo ótico e vetores de movimento MPEG-4, para obter as informações sobre o movimento que serão depois aplicadas no cálculo da homografia. De seguida, comparamos os resultados desta metodologia com técnicas do estado da arte para avaliar a sua eficácia e potenciais melhorias.

Os métodos propostos mostraram-se promissores, integrando os dados de movimento diretamente nas equações de homografia bem como na geração dos panoramas. Embora menos precisos do que os métodos mais avançados, as nossas abordagens demonstraram um erro reduzido em imagens com poucos pontos de interesse. Tanto o fluxo ótico como os vetores de movimento MPEG produziram resultados semelhantes. As observações qualitativas dos resultados da super imagem forneceram informações relevantes, proporcionando uma perspetiva única em comparação com outros métodos. As nossas conclusões sugerem que a incorporação de dados de movimento pode ser utilizada para melhorar a estimativa da homografia, particularmente em cenários com poucos pontos de interesse.

Abstract

Football is a highly competitive activity, where the slightest difference in performance can significantly impact a game's outcome. Due to its worldwide popularity, it drives many investments to reach success. Consequently, teams are compelled to use this financing to acquire better results. As the challenges grow in both economic and sporting domains, technologies are emerging as a solution to optimise performance and gain an edge over rivals.

Automatic video analysis is one of the key technologies currently used to optimise strategies. However, most existing solutions require expensive and complex hardware, such as multiple fixed cameras, sensors, or GPS devices. This poses a significant challenge for many clubs, especially those with limited resources. Regarding the adversities the previous solutions may raise, using images and video in scenarios with minimal restrictions, particularly with a moving camera, allows a broader audience to benefit from cutting-edge technology. However, a moving camera introduces different challenges due to the varying images that are caused by motion. Therefore, the algorithms for video analysis should be reliable and adaptable to handle these challenges.

This research proposes an approach for homography computation in football matches using moving cameras, addressing the limitations of current methods in dynamic and complex environments. Existing techniques, which rely heavily on visible keypoints and are applied frame-by-frame, often lack temporal consistency and struggle in areas with few keypoints. To address these challenges, we propose two primary methods: creating a superimage panorama using camera motion and integrating motion data directly into homography equations. We leverage optical flow and MPEG-4 motion vectors to utilise their motion information for homography computation. We will then compare the results of this methodology with state-of-the-art techniques to evaluate its effectiveness and potential improvements.

Our proposed methods show promise by integrating motion data directly into homography equations and generating panoramas. Although less accurate overall than state-of-the-art methods, our approaches demonstrated reduced error in keypoint-sparse frames. Both optical flow and MPEG motion vectors produced similar results, with MPEG proving more efficient due to reduced processing. Qualitative observations of panoramic outputs delivered valuable insights, providing a unique perspective compared to other methods. Our findings suggest that incorporating motion data can be used to enhance homography estimation, particularly in challenging conditions with fewer keypoints.

Sustainable Development Objectives

In 2015, the Member States of the United Nations (UN) established "The 17 Sustainable Development Goals" (SDGs). They delineate a significant plan for achieving a better and more sustainable future for all, underlying environmental, social and economic aspects. In context with these goals, the importance of innovation, infrastructure, and reduced inequalities play an essential role, which is particularly relevant in sports technology.

This work focuses on the sophisticated world of sports analysis, concentrating on the role of automatic video processing in enhancing team performance and strategic planning. In this context, understanding how Sustainable Development Objectives (SDOs) can be integrated into sports technology is crucial.

This research is driven by the motivation to bridge the gap between high-end technological solutions and their accessibility to smaller clubs and competitions. In simpler terms, to make these advanced technological solutions available to all competitive entities, even if they cannot afford the typical costs associated with existing advancements. Through the leveraging of affordable and efficient video analysis tools, even teams with limited resources can enhance their performance and strategic capabilities. This aligns with the more general aim of ensuring that the benefits of technological innovation are equitably distributed, thereby supporting the SDOs.

Beyond technology, the principles of fair play, inclusivity, and community engagement are crucial to the sustainable development of sports. Moreover, the high-cost advanced systems usually require substantial computational power units, which in turn consume considerable amounts of energy. In light of this, leveraging low-restriction scenarios, especially those utilising a single moving camera, can offer a more sustainable alternative.

Acknowledgements

First, I want to thank my supervisor, Professor Pedro Carvalho, for all the help and guidance along the development of this work.

To my family, especially my parents, thank you for your persistent support and patience throughout this journey. I am grateful for your constant motivation, which has been my pillar of strength.

Finally, I want to extend my appreciation to all my friends for the laughter, shared memories, and companionship over the past five years. Your kindness made everything seem easier than it truly was, and I am genuinely grateful for your support.

Carlos Gomes

"The more difficult the victory,
the greater the happiness in winning."

Pelé

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 2
- 1.3 Objectives . . . 2
- 1.4 Document Structure . . . 3

2 Background **5**
- 2.1 The anatomy of the field of play . . . 5
- 2.2 Machine Learning & Deep Learning . . . 6
- 2.3 Neural Networks . . . 7
- 2.4 Computer Vision . . . 9
  - 2.4.1 Edge Detection . . . 9
  - 2.4.2 Segmentation . . . 10
  - 2.4.3 Homography . . . 10
  - 2.4.4 Image Stitching . . . 12
  - 2.4.5 Optical Flow . . . 13
  - 2.4.6 Codec MPEG . . . 13

3 State of the Art Review **15**
- 3.1 Technology Review . . . 15
  - 3.1.1 AI Supported Video Assistant Referee Systems . . . 16
  - 3.1.2 Automated Broadcast Cameras . . . 17
  - 3.1.3 Mediacoach and Beyond Stats . . . 18
  - 3.1.4 Track160 . . . 19
  - 3.1.5 Genius Sports and Second Spectrum . . . 20
  - 3.1.6 Stats Perform . . . 21
- 3.2 Emerging Trends in Football Video Analysis . . . 22
  - 3.2.1 Computer Vision in Sports . . . 22
  - 3.2.2 Calibration & Homography Calculation . . . 24
- 3.3 Relevant Datasets . . . 35
  - 3.3.1 IPL Azadi Soccer Dataset . . . 35
  - 3.3.2 SoccerNet Dataset . . . 35
  - 3.3.3 World Cup 2014 Dataset . . . 36
  - 3.3.4 TS-WorldCup Dataset . . . 37
  - 3.3.5 SportsMOT Dataset . . . 37

4 Research Proposal **39**
4.1 Problem Definition . . . 39
4.2 Proposed Solution . . . 40

5 Methodology **43**
5.1 Implementation Strategy . . . 43
5.1.1 Automatic State-of-the-Art Methods . . . 44
5.1.2 Manual Homography Calculation . . . 45
5.1.3 Motion Calculation . . . 45
5.1.4 Super Image Creation . . . 51
5.1.5 Directly Applying the Motion Information Into the Homography Equations 54
5.2 Metrics for Evaluating Homography Estimation . . . 55
5.2.1 Intersection-over-Union . . . 55
5.2.2 Distance between players' positions . . . 56
5.3 Data and Hardware Overview . . . 57

6 Results **59**
6.1 State-of-the-Art Methods Evaluation . . . 59
6.2 Super Image Approach . . . 62
6.3 Equations Approach . . . 67

7 Final Discussion **71**
7.1 Conclusions . . . 71
7.2 Future Work . . . 73

References **75**

A Panoramas **85**
A.1 Video 1 Results . . . 85
A.2 Video 2 Results . . . 86
A.3 Video 4 Results . . . 86
A.4 Video 5 Results . . . 87

List of Figures

2.1 Field of play dimensions and pitch markings . . . 6
2.2 Relation between Artificial Intelligence, Machine Learning and Deep Learning . 7
2.3 Neural Network Example . . . 8
2.4 Example of homography application . . . 11

3.1 FIFA Semi-Automated VAR . . . 16
3.2 Mediacoach and Beyond Stats . . . 19
3.3 Track 160 Applications . . . 19
3.4 Genius Sports and Second Spectrum . . . 20
3.5 Opta Vision Use Cases . . . 22
3.6 AI Role in Sport . . . 23
3.7 DL Techniques in Sports Application . . . 24
3.8 Football Field Reference Markers . . . 26
3.9 Pan, Tilt, Zoom Camera and Real World Coordinate System . . . 28
3.10 The Six Straight Lines used as Features . . . 33
3.11 SoccerNet Dataset Data . . . 36
3.12 WordlCup 2014 Dataset Examples . . . 37
3.13 TS-WorldCup Data . . . 38
3.14 SportsMOT Data . . . 38

4.1 Challenges faced for homography calculation in football . . . 40
4.2 Proposed approach using the panorama . . . 41

5.1 Pixel coordinate system . . . 46
5.2 Relation between X-axis displacement and camera motion . . . 47
5.3 Comparison of the original frame and the visualisation of the optical flow . . . 49
5.4 Comparison of the original frame and the visualisation of the motion vectors extracted from MPEG encoding . . . 50
5.5 Concatenation Example . . . 51
5.6 Annotated points example and respective generated mask . . . 56

6.1 MSE Histogram - TVCalib Video 4 . . . 60
6.2 Examples of TVCalib Outliers . . . 60
6.3 MSE Histogram - Sportsfield Video 4 . . . 61
6.4 MSE Histogram - Sportsfield Video 2 . . . 62
6.5 Examples of Sportsfield Outliers . . . 62
6.6 Evolution of IoU according to the distance relative to the first frame - Video 1 . . 63
6.7 MSE Results of Video 4 Using the Super Image Approach . . . 64
6.8 Comparison of Predicted Players' Positions for an Outlier in Video 5 . . . 65

6.9 Panorama Video 3 - Optical Flow 65
6.10 Panorama Video 3 - MPEG 66
6.11 Original Frame 66
6.12 Example of a Frame mapped in a Super Image generated using MPEG motion vectors 67
6.13 Evolution of IoU over the video frames 68
6.14 Comparison of Predicted Players' Positions for an Outlier in Video 5 68
6.15 MSE Results of Video 4 Using the Super Image Approach 69

A.1 Panorama Video 1 - Manual Motion 85
A.2 Panorama Video 1 - Optical Flow 85
A.3 Panorama Video 1 - MPEG 85
A.4 Panorama Video 2 - Optical Flow 86
A.5 Panorama Video 2 - MPEG 86
A.6 Panorama Video 4 - Optical Flow 86
A.7 Panorama Video 4 - MPEG 87
A.8 Panorama Video 5 - Optical Flow 87
A.9 Panorama Video 5 - MPEG 87

List of Tables

5.1 Video's Description 58

6.1 IoU Results of the State-of-the-Art methods 59
6.2 IoU Results of Super Image Approach 63
6.3 IoU Results of Equations Approach 67

Abbreviations and Symbols

2D	Two Dimensions / Two Dimensional
3D	Three Dimensions / Three Dimensional
AI	Artificial Intelligence
ATP	Association of Tennis Professionals
CNN	Convolutional Neural Networks
CV	Computer Vision
DFL	*Deutsche Fußball Liga* (Germany's Football League)
DL	Deep Learning
DLT	Direct Linear Transform
EA	Electronic Arts
FIFA	*Fédération Internationale de Football Association*
FPS	Frames per Second
GAN	Generative Adversarial Networks
HSV	Hue, Saturation, Value colour space
Hz	Hertz
IFAB	International Football Association Board
IoU	Intersection-over-Union
MKV	Matroska Video
ML	Machine Learning
MSE	Mean Squared Error
NN	Neural Networks
PTZ	Pan, Tilt, and Zoom
RAFT	Recurrent All-Pairs Field Transforms
RGB	Red, Green, Blue colour space
ReLU	Rectified Linear Unit
SDOs	Sustainable Development Objectives
Stats	Statistics
Std	Standard Deviation
Tanh	Hyperbolic tangent
UEFA	Union of European Football Associations
UN	United Nations
VAR	Video Assistant Referee
WTA	Women's Tennis Association
XOR	Exclusive OR

Chapter 1

Introduction

Wide-spectrum sports, with a particular emphasis on ball-based disciplines, involve multiple movements, positioning and interactions between players and the ball. In an increasingly competitive environment, understanding these aspects is crucial to improving the performance of the teams and their individuals. However, analysing football games is challenging since finding these patterns takes much time, especially in moving camera scenarios. Therefore, coaches, players, and analysts must use video analysis processed by algorithms that can handle such systems.

1.1 Context

Automatic video processing is emerging as a solution to cope with the growing competition in sports that demands efficiency and innovation from the teams, since the role of football analysis can impact team performance in many ways. By examining games and training sessions through the lens of player positions, trajectories, actions, and gameplay, clubs can enhance both their performance and processes.

Team analysts typically conduct this game analysis on performance and statistics and subsequently communicate it to the managers. Clubs must also study future opponents to plan the strategy for the incoming matches to identify the virtues and weaknesses of rivals and take advantage of them. Video processing can also impact transfer decisions by evaluating potential signings. However, this analysis is time-consuming as it requires watching for many hours to find behaviours in teams' gameplay. Therefore, more than a single match analysis is required to gain valuable insights into game patterns, making automatic video processing a growing trend in sports. In addition, automated video processing is also used for other purposes. It can assist referees in their decisions with accurate and fair calls, such as determining offside decisions, which is challenging without these technologies. It can also benefit entities not directly involved in the game, for instance, fans, sponsors and journalists. Some examples are using three-dimensional (3D) reconstruction to create engaging content for fans or to insert virtual ads into live matches, automatically detect highlights, and create live commentary using artificial intelligence.

1.2 Motivation

Football is a highly competitive activity, where even the smallest performance nuances can dramatically alter the outcome of a game. Its worldwide popularity, coupled with its escalating competitive landscape, drives substantial investments into clubs. As a result, teams are driven to strategically utilise these funds to enhance their performance, ultimately aiming for sportive success. With this financing, clubs are more capable of hiring more elite professionals. However, as previously noted, analysing games is a time-consuming process. The advent and advancements in Artificial Intelligence (AI) and technology have emerged as a game-changer in the contemporary sports landscape, making game analysis more efficient. Top-tier teams playing in prestigious competitions draw a vast number of spectators, attracting a larger pool of investors and sponsors and increasing the prize money. This increased financial backing, in turn, provides the means for these teams to invest in transformative technological solutions. Even though these innovative solutions have the potential to improve a team outcome, they are usually costly and require expensive equipment and hardware, which makes the accessibility and affordability of these advanced solutions a challenge for many. Thus, while technology presents immense potential, its widespread adoption across all clubs and competitions is yet to be realised, thereby missing out on opportunities to enhance performance.

In this context, using images and video in unconstrained scenarios, especially with a moving camera, allows even smaller clubs and competitions to benefit from cutting-edge technology. This approach enhances their performance analysis and player development without incurring the costs associated with other high-tech solutions. This method is of great interest due to its ability to capture and represent the complex spatio-temporal characteristics of the objects and events under observation with exceptional precision and detail.

The non-intrusive nature of relying solely on images and videos captured by a moving camera eliminates the need for additional equipment or sensors, which could potentially interfere with the players' behaviour or performance, making it even more cost-effective. As a result, it enables a more genuine and reliable observation of the game and its unfolding dynamics, providing a true-to-life representation of the sport. This ensures that the analysis is grounded in reality, offering valuable insights for performance enhancement.

1.3 Objectives

This research is dedicated to homography calculation in football matches, particularly in moving camera scenarios. The main purpose is to use efficient and easily accessible video processing techniques for sports analysis.

We aim to analyse errors introduced when insufficient keypoints are available for accurate homography calculation. Understanding these limitations will potentially help to identify strategies to mitigate these issues and improve robustness of current techniques.

Moreover, we will explore how information from camera movements can effectively, approximate the homography transformation. By testing motion data integration, we aim to determine if it can enhance homography calculations, even in dynamic and complex football match environments.

Through these efforts, we aim to analyse how different motion-based homography calculation methods perform comprehensively. By evaluating and comparing these techniques, we strive to determine their effectiveness and potential improvements for homography calculation in football matches, particularly in moving camera scenarios.

1.4 Document Structure

Apart from the introduction, this document contains six more chapters. Chapter 2 describes some topics related to the theme and covered in the state-of-the-art chapter. Chapter 3 describes the state-of-the-art and presents related work and some of the available commercial solutions. Chapter 4 provides an overview of the problem and our planned solution to achieve the research goals. Chapter 5 explains the methodology and experimental setup employed in our study. It encompasses a comprehensive overview of the methodologies employed, a detailed discussion of the datasets used, the metrics applied for evaluation, and the strategic approach adopted for conducting the experiments. Chapter 6 analyses the outcomes derived from the experimental setup. It includes a comparative study of the methods employed, their effectiveness and performance in the context of our research. Chapter 7 serves as a reflective assessment of our research, offering a conclusive summary of the study, evaluating the accomplishment of the research goals, and outlining potential directions for future work.

Chapter 2

Background

In this chapter, we will explore and explain some concepts mentioned in the following sections and chapters. This chapter is structured to provide a guide to the topic, beginning with an introduction to key concepts that form the foundation of this field. This knowledge covers the essential principles and methodologies that aim to provide understanding of how computer vision is applied in analysing football games and is crucial for grasping more complex details presented afterwards.

2.1 The anatomy of the field of play

In the camera calibration process, especially within the dynamic environment of football matches, a comprehensive understanding of the field's dimensions and anatomy is not just advantageous—it is indispensable. We will explore the standardised measurements that delineate a football field's length, width, and area, along with various marked regions such as the penalty area, goal area, and centre circle. Learning these measurements is crucial for establishing reference points and coordinate systems in camera calibration. Furthermore, we will delve into the importance of features like goalposts and corner flags, which frequently act as essential reference markers in calibration algorithms. This knowledge plays a pivotal role in enhancing the precision and effectiveness of video analysis in football.

The playing field of a football match is a rectangular space demarcated by continuous lines that define the boundaries of specific areas. The more extended boundaries of the rectangle are known as touchlines, while the shorter ones are referred to as goal lines. It is a rule that the touchlines must always be longer than the goal lines.

The Fédération Internationale de Football Association (FIFA) recommends a field of play dimension of 105 metres by 68 metres [30]. However, these dimensions are not strictly enforced. In fact, there is a range within which competitions may determine the length of these lines. For national-level matches, the dimensions offer some flexibility. The length of the pitch can vary from 90 meters to 120 meters, and the width from 45 meters to 90 meters (100 yards). This range accommodates different stadium sizes and local conditions. For international matches, the dimensions are more specific. The standard pitch size is 105 meters long and 68 meters wide. This

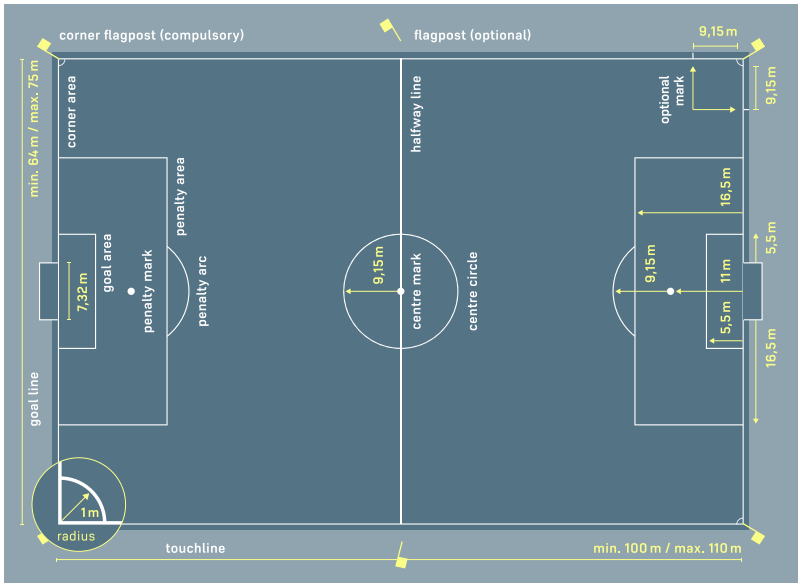


Figure 2.1: Field of play dimensions and pitch markings (Extracted from IFAB (2023) [6])

uniformity ensures that international matches are played under consistent conditions, providing a level playing field for all teams [6].

The football field is bisected by a halfway line, which connects the midpoints of the touchlines, effectively dividing the field into two equal halves. A centre circle is distinctly marked around the centre mark, serving as a boundary during kick-offs.

In front of each goal, the goal area and penalty area are clearly demarcated. Additionally, corner areas, represented by quarter circles, are outlined within the field of play near the four corners of the pitch. The specific dimensions of these areas are illustrated in Figure 2.1. It is important to note that apart from the length and width of the field, which may sometimes vary slightly, every other measurement has a fixed and mandatory size, as indicated in Figure 2.1. This includes the goal area and penalty area lines' lengths, the centre circle radius, corner areas, respective flagposts, and the goal itself.

Each goal line houses a centrally positioned goal composed of two vertical posts connected at the top by a horizontal crossbar. The distance between the posts is 7.32 meters, and the height from the ground to the crossbar is 2.44 meters [5].

2.2 Machine Learning & Deep Learning

Machine Learning (ML) is a category of AI that uses data and algorithms to learn from experience and make decisions. In other words, it allows computers to forecast based on historical data [40]. Machine learning is a field that is divided into four main types: Supervised Learning is used for classification, regression, and time-series forecasting tasks; Unsupervised Learning is associated with functions like clustering and anomaly detection; Semi-supervised Learning is practical when labels are limited and include use cases from supervised and unsupervised Learning, such as active

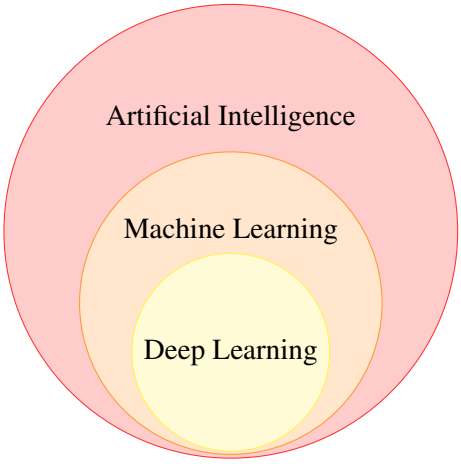


Figure 2.2: Relation between Artificial Intelligence, Machine Learning and Deep Learning

Learning; Reinforcement Learning is commonly related with game playing and robotics, but it is beneficial for several real-life scenarios and applications, enabling systems to make informed decisions and take actions in dynamic environments. Each type has its unique approach to learning from data and making predictions or decisions. There isn't a single algorithm that works well in every circumstance or problem [58].

Deep Learning (DL), on the other hand, is a subset of Machine Learning that uses neural networks with many layers to learn from complex data. Neural networks are composed of artificial neurons that process and transmit information. Deep Learning can handle unstructured data, such as images and text, and automatically learn features. It has advanced computer vision, natural language processing, and speech recognition [39] [35]. Figure 2.2 illustrates the relation between Artificial Intelligence, Machine Learning and Deep Learning. In broad terms, AI spans the spectrum of creating and exploring intelligent machines. In contrast, ML empowers machines to glean data insights and autonomously make decisions. Deep Learning, as its name suggests, focuses on specialised algorithms that closely mimic the intricacies of the human brain. This specialisation endows computers with the capacity to discern intricate patterns from vast datasets, akin to the way biological learning occurs [44] [80].

2.3 Neural Networks

Neural Networks (NN), often called Artificial Neural Networks, are learning and optimisation algorithms based on neuroscience theories [62]. Instead of relying upon predefined algorithms or manually created features, neural networks learn from examples to find solutions to problems by automatically creating features on raw data. This learning-based approach is inspired by the information-processing mechanisms in biological nervous systems, particularly the human brain [9]. A neural network functions as a nonlinear mathematical system, transforming input variables into output variables. This transformation is governed by parameters known as weights, which are adjusted or "learned" through exposure to a set of examples. This learning or training process

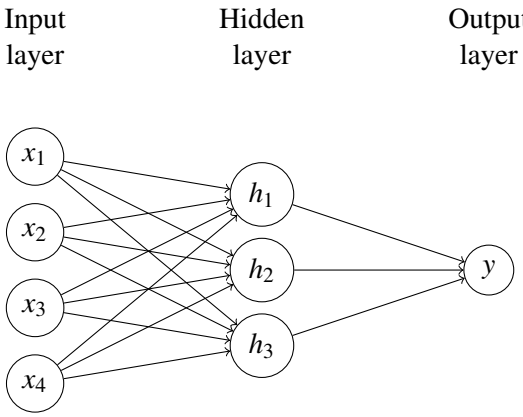


Figure 2.3: Neural Network Example (Adapted from Alemu et al.[1])

can be computationally intensive. However, once the network is trained, it can process new data rapidly. A fundamental feature of neural networks is the capability to induce specific examples to broader solutions, often circumventing the need for a first-principles model of the underlying processes. Challenges arise from the need for appropriate training data and the potential issues when networks extrapolate beyond the scope of their training data. However, various techniques can mitigate these issues.

The example neural network in Figure 2.3 depicts a simple feedforward neural network. The network begins with the input layer, where each node represents one input data feature. The image has four input nodes, meaning the network accepts a four-dimensional input vector. Each feature in the input data is fed into its respective input node. Each edge (line) connecting nodes from the input layer to the hidden layer represents a weight. These weights are parameters that the network will learn during training. It determines how much influence the input node will have on the hidden node it is connected to. The hidden layer processes the inputs received from the input layer. Each node in this layer takes the sum of the weights of its inputs and applies an activation function. The activation function is a mathematical function that introduces non-linearity into the network, allowing it to learn complex patterns. Standard activation functions include sigmoid, rectified linear unit (ReLU), and hyperbolic tangent (Tanh). After the hidden layer nodes have processed the inputs, they are again connected to the output layer via weighted edges. The output layer in this network consists of a single node, which suggests that the network is designed to perform a task like binary classification or regression with a single output variable. It applies another activation function to produce the final output of the network. The network uses a dataset with known inputs and outputs to adjust the weights during training. The goal is to minimise the difference between the predicted output and the actual value in the training data. This process is called backpropagation, where the error is calculated at the output and propagated back through the network, updating the weights along the way [44] [3].

Regarding computer vision, Convolution Neural Networks (CNN) are extensively used in image and video analysis as they automatically extract information from unprocessed input photos [69]. Even though its computational complexity affected its application, the exponential develop-

ment of technology, including hardware, allowed the application in larger scales [16] such as face recognition, autonomous vehicles, self-service supermarkets, and intelligent medical treatments [47]. The main distinction from classic neural networks is the convolution layer, which performs feature extraction by applying the convolution operation. This is done by using a filter, also known as kernel, to an input image. It then generates a feature map with features like edges, lines or corners. This map is fed into an activation function before going to the next layer. The next layer is usually one of the following: the pooling layer, whose goal is to downsample the feature map. Maximum, minimum, and average are the most common operators to lower the dimensionality. To prevent overfitting, batch normalisation, which, as the name explains, normalises the output of the previous layer, is commonly used. More than regularising overfitting, it speeds up the training using a higher learning rate. Dropout is also used as a mechanism for overfitting prevention. It randomly removes neurons during each iteration, which slightly modifies the network. A CNN architecture usually ends with at least one fully connected layer to generate the final output. Here, each neuron from the previous layer is connected to each neuron on the current layer. This complete linkage gives the name to this type of layer. It has the same number of nodes as the output classes and generates the output based on the acquired features [69].

2.4 Computer Vision

Computer Vision (CV) is the scientific and technological domain that enables machines to see and understand the visual world. It is a technology that uses Artificial Intelligence (AI) to allow computers to perform tasks that humans can do with their eyes and brains, such as recognising objects, faces, scenes, actions, and gestures, reconstructing three-dimensional (3D) models from images and videos, enhancing and restoring content, and creating realistic visual effects.

The history of computer vision can be traced back to the 1960s, when researchers started to explore the possibility of using computers to analyse images. Some early achievements include finding edges, corners, and lines in images, segmenting images into regions, and recognising simple shapes and patterns.

The current decade has seen remarkable progress in computer vision, driven by the availability of large-scale data sets, deep Learning, and mobile devices. Computer vision has achieved impressive results in many tasks, such as image classification, semantic segmentation, face verification, human pose estimation, scene understanding, and image captioning. Computer vision has also become more accessible and ubiquitous, as it is integrated into various products and services, such as social media, e-commerce, gaming, health care, and autonomous driving [90].

2.4.1 Edge Detection

Edge detection is fundamental in image processing and is used to identify points where the brightness or colour of the image changes significantly. This typically indicates a boundary between different regions in the picture, such as the edge of an object. Edge detection is crucial in many applications, including image analysis, pattern recognition, and computer graphics [90] [68]. At

its core, edge detection involves examining the local neighbourhood of each pixel and determining whether there is a change in intensity or colour that signifies an edge. This change is often measured by calculating, at each pixel, the gradient of the image intensity.

Edge detection provides real challenges in terms of noise sensitivity, as noise can create false edges. In digital imaging, noise refers to random variations of brightness or colour information, which may have many causes, such as lighting conditions or sensor limitations. This noise can lead to false edges, where the algorithms incorrectly identify noise variations as significant changes in image intensity or colour, thus mistakenly detecting edges that do not correspond to the actual boundaries of the object in the image. Determining the appropriate thickness of an edge also presents challenges. An edge detection algorithm should identify the edges with precise, thin lines. However, in practice, edges might be detected as thicker bands rather than sharp lines, leading to ambiguity about the exact location of the edge.

2.4.2 Segmentation

Image segmentation is a widely researched area where the objective is to partition an image into meaningful segments or groups of pixels that share specific characteristics. This process is analogous to clustering in statistics, where data points are grouped according to similar features. Image segmentation algorithms aim to identify these groups in an image, effectively splitting regions that define distinct objects, parts or groups of objects [90].

In image segmentation, various techniques divide images into meaningful segments. Each method provides a unique approach to segmenting images into distinct, significant parts. Moreover, to enhance the segmentation results, we can apply some morphological operations, as they help manipulate the shape and structure of binary images, making them more suitable for segmenting objects of interest from the background. Morphological operations are a set of methods designed for binary images, where each pixel can have one of two values: 1 (white) or 0 (black) [90]. There are four common operations. Firstly, dilation adds pixels to the borders of objects in a binary image, expanding the shapes. This operation helps join nearby objects and fill small holes. The opposite of dilation, erosion removes pixels on object boundaries. Erosion helps remove slight white noise, detaching two connected objects, or thinning objects. This operation is a combination of erosion followed by dilation. Opening an image removes small objects and protrusions. It is beneficial for eliminating noise. Closing is the reverse of opening, involving dilation followed by erosion. It helps close holes inside foreground objects or minor black points on the object.

2.4.3 Homography

Homography is a fundamental concept in the fields of Vision and geometry, particularly in the context of image processing. It describes a transformation between two planes in a 3D space, and it is used in tasks like image rectification or registration, perspective correction, and camera calibration [21]. This transformation is crucial in understanding how the 3D world is projected

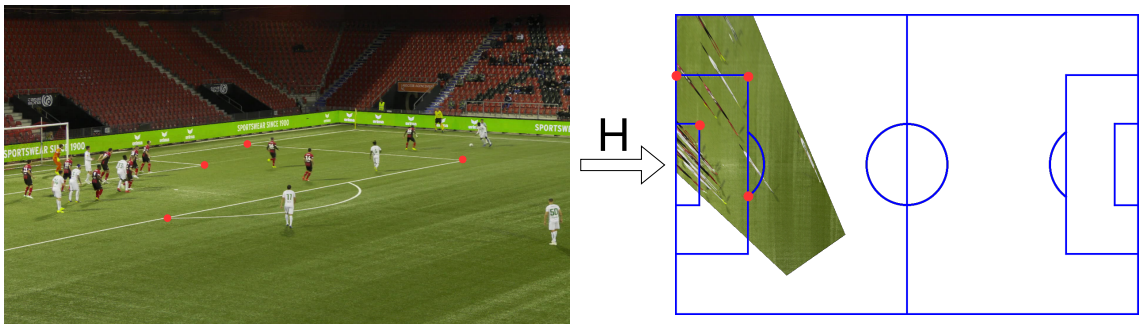


Figure 2.4: Example of homography application

onto 2D images through cameras. In other words, understand the relationship between points in the real world and their projections in images.

Theoretically, homography is a projective transformation between two planes or a mapping between two planar projections of an image. It represents the most straightforward transformation that describes the 2D relationship between two images, particularly when the camera or the observed object moves [74]. This concept is rooted in projective geometry, a mathematical tool for describing multi-view geometry.

Homographies are represented in a homogeneous coordinate system, allowing for more complex transformations than standard 2D coordinates. A 3D conversion in this coordinate space can mathematically describe a homography. The change itself is represented by a 3x3 matrix known as the homography matrix (H). When multiplied by a point in one plane, this matrix gives the correspondence on the other plane.

Figure 2.4 presents a visual representation of how it works. The image shows a broadcast frame of a football match on the left. The arrow denotes the homography matrix, which represents the transformation that maps points in the broadcast frame to their corresponding positions in the field model. By applying the homography matrix to the entire image, the broadcast frame is accurately warped to fit the field template, making it possible to see which part of the field is seen in the frame.

This transformation between two planes relates as follows:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \mathbf{H} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2.1}$$

where $[x',y',1]$ represents the homogeneous coordinates of a point in the second image (or destination image). The matrix H is the homography matrix. It is a 3x3 matrix that encapsulates the geometric transformation between the two images. Each element h_{ij} in the matrix corresponds to a specific aspect of the transformation. Lastly, $[(x,y,1)]$ the homogeneous coordinates of the corresponding point in the original image [55]. So, using Figure 2.4 as an example the marked points

on the frame would represent the [(x,y,1)] points, and their respective, marked on the template, would define the [x',y',1], obtained by applying H.

Despite having nine elements, the matrix has only 8 degrees of freedom because scaling the matrix by a non-zero constant doesn't change the transformation, a characteristic feature of homogeneous coordinates[21]. Because of that, a minimum of four-point correspondences are needed to compute the homography. Each point provides two equations, for x and y, so to get eight equations, we need at least four pairs of points[55].

An essential property of homographies is invertibility. This property is an important characteristic that allows the transformation to be reversed or undone, which is crucial in many applications like 3D reconstruction or calibration. Another feature of homographies is that they preserve the collinearity of points. If three points are on a straight line in one plane, their transformed counterparts will be on a straight line in the other plane. Each homography can be decomposed into an affine transformation, which includes rotation, translation, scaling, and shearing, and a metric (or Euclidean) transformation, which provides for only rotation and translation [74].

However, this estimation process could be more straightforward due to noise and inaccuracies in real-world data. In an ideal scenario without noise, the estimation of a homography can be solved with a linear system using the Direct Linear Transform (DLT) algorithm. However, in reality, this becomes an optimisation problem, often solved using statistical methods which assume that the image points are random variables with Gaussian distributions and aim to minimise the reprojection error. This is where the DLT algorithm is typically combined with a robust estimation method like RANSAC (Random Sample Consensus). RANSAC is designed to handle outliers and non-Gaussian noise effectively, making it crucial for practical applications. It randomly selects a subset of the point correspondences, applies the DLT algorithm to estimate the homography, and then determines how well the estimated homography fits the entire set of point correspondences. This process is repeated multiple times, and the homography that results in the minimum reprojection error is selected as the final estimate. In cases where the assumptions of Gaussian noise do not hold, these methods are particularly valuable [21].

2.4.4 Image Stitching

Image stitching is a computer vision task in which the objective is to combine multiple images captured from different points of view but with overlapping. The goal is to create a high-resolution panoramic image encompassing all the views into a single, wide, cohesive composition. This process, also known as mosaicing, finds applications in various fields. Medical applications, satellite imaging, industry, and image reconstruction are some of the examples. [23].

The typical workflow in image stitching begins with keypoint detection and feature matching. This enables the creation of correspondences between similar points across multiple images. Once a robust set of correspondences is established, they are utilised to compute the homography matrix between pairs of images. As previously stated, this matrix will allow us to define the transformation between two image planes anchored by the detected and matched keypoints. Outlier removal techniques may be employed to enhance the precision of this transformation. Finally, with the

homography matrix determined, an image is warped by applying the matrix, thereby converting one image's coordinate system with another. To create a seamless composition, we must blend the warped images. Various blending techniques are available to ensure a smooth transition and visually appealing panoramic composition. [102] [33] [53].

While image stitching offers numerous advantages and applications in various fields, it has challenges. These include struggling with significant parallax situations and real-time dynamic changes, which can lead to mismatching keypoints. Additionally, processing extensive collections of images can be time-consuming as each image requires individual attention. Another notable challenge is the elimination of visible seams in the final stitched image.

2.4.5 Optical Flow

Optical flow is characterised as the pattern formed by the perceived motion of objects between two consecutive frames. It is a 2D vector field, where each vector represents the displacement. This displacement illustrates the movement of points from the first frame to the second frame along both axes [66]. It represents the projection of the three-dimensional motion of objects onto an image plane. This can be manifested as either instantaneous velocities or discrete displacements. This concept is fundamental in understanding the dynamics of visual scenes and the interaction of objects within them [7].

Optical flow works on some assumptions: the intensity of an object's pixel remains unchanged between consecutive frames, and the spatial coherence suggests that neighbouring pixels exhibit similar motion [66]. These assumptions are crucial for the accurate estimation of motion. However, the process is further complicated by challenges such as occlusion, transparency and the aperture problem.

In this context, dense optical flow stands out as a computer vision technique that estimates objects' motion across consecutive frames. In contrast to sparse optical flow, which calculates flow vectors for select features considered as interesting, dense optical flow demonstrates flow vectors for every pixel within the frame. Although dense methods are generally slower, they have the potential to yield more precise results [100].

Among the main applications of this technology are motion estimation, which involves calculating the movement of objects within a frame; video compression, which reduces file size by encoding only the changes between frames; image interpolation, which generates intermediate frames to create smooth motion in videos, and video stabilisation, which corrects for camera shake to produce smoother footage. Additionally, semantic segmentation, object detection, and tracking are critical use cases that enable detailed scene analysis and object monitoring in various contexts.

2.4.6 Codec MPEG

A codec is an abbreviation for coder and decoder applied to signals or data streams. They are used to decompress or compress videos to play them. The 1980s saw the rise of digital video, allowing for digital video compression in various applications. The Moving Picture Experts Group

(MPEG)[1] is responsible for developing standards for coded representation of moving pictures, sounds, and their combination. MPEG codec enables efficient transmission and storage of high-quality video and audio content while preserving the perceptual quality and compatibility with different devices and platforms [4].

Regarding that compression ratio, MPEG relies on the reduction of spatial and temporal redundancy. MPEG uses techniques like the Discrete Cosine Transform (DCT) to convert these spatially redundant areas into a format that can be more efficiently compressed. DCT helps to identify parts of the image that do not change much and compress them more, thus reducing the amount of data needed to represent a single frame [4]. In a typical video, many frames are similar to each other, especially in sequences where there is little movement. MPEG identifies and eliminates these redundancies by storing only information about the differences between successive frames rather than storing each entirely [24]. This is achieved through techniques like motion estimation and compensation. In most cases, MPEG only stores the changes (motion vectors and difference data) from one frame to the next rather than storing each entire frame.

[1] https://www.mpeg.org/

Chapter 3

State of the Art Review

In the following chapter of this document, we will explore the current landscape in the field of football game analysis using computer vision.

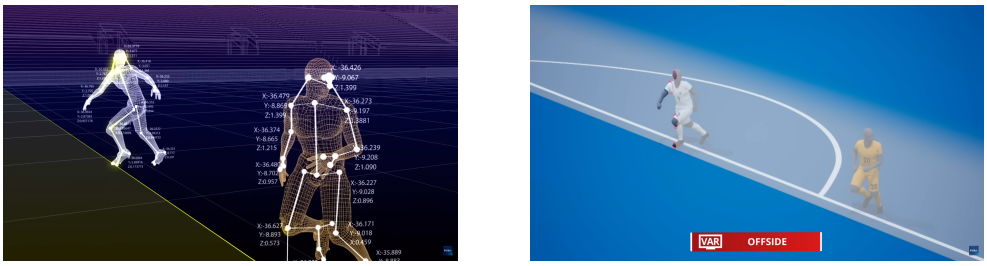
This chapter is structured to provide a detailed and layered understanding of the topic. Beginning with a conceptual introduction, we delve into a review of existing technologies that are already being utilised in the market. This section will highlight the practical applications of computer vision in football game analysis, showcasing how professionals in the industry are currently leveraging these technologies.

Next, we transition to an analysis of related work. This part of the chapter will focus on recent academic studies offering insights into the cutting-edge developments and emerging trends in this domain. We investigate camera calibration and homography estimation methods and how they are applied.

Finally, the chapter concludes by discussing relevant datasets instrumental to our work. These datasets are crucial in developing and validating computer vision technologies in football game analysis.

3.1 Technology Review

The game we know today as football was formalised in England in the 19th century and later spread globally, thanks to the Olympic Games, the *FIFA World Cup*, and professional leagues [10]. Since then, many technological innovations and developments have emerged to enhance the sport's performance, analysis, and enjoyment. These technologies range from hardware devices, such as cameras, sensors, and wearables, to software applications for multiple purposes, such as data analysis, betting, and virtual reality. In this section, we will see through the existing technologies being used for football and discuss the challenges and opportunities they present for the industry.



(a) Detection of 29 data points of each individual (b) 3D animation generated by the system

Figure 3.1: FIFA Semi-Automated VAR (Extracted from FIFA, 2022[26])

3.1.1 AI Supported Video Assistant Referee Systems

Video Assistant Referee, also known as VAR, is a system that helps refereeing teams review crucial decisions such as goals, penalties, red cards, and mistaken identity. It was introduced to deal with human error and bias, ensuring that more accurate and fair calls are made.

FIFA has always been trying to improve the game with technology. One of the latest innovations was the Semi-Automated Offside Assistant used in the 2022 World Cup in Qatar. It is a video tool to support officials to make faster and more accurate offside calls. It needs 12 tracking cameras mounted on the stadium roof and a sensor placed inside the ball. This sensor, an inertial measurement unit, sends data to the VAR room at 500 Hertz(Hz), making the touch detection very precise. Then, the cameras detect each individual's ball and 29 data points, including all limbs and extremities relevant to these situations (Figure 3.1a). By combining ball tracking and those data points, the system applies artificial intelligence to alert video referees to an offside alert, which analyses by manually checking the selected kick point and offside line previously created using the limb points before communicating it to the on-field referees. After the decision is made, the system also generates a 3D animation (Figure 3.1b) that perfectly details the position of the players at the moment the ball was played using the same 29 data points shown on the stadium's big screen and in the TV broadcast [29] [28]. The generation of 3D animation complements the improvement of the accuracy in decision-making with spectator experience by providing explicit visual representations of these decisions. As we continue to explore and innovate in this field, it is crucial to ensure that these technologies are designed and implemented to promote clear communication and engagement with the spectators. This will help foster a better understanding and acceptance of AI-assisted decision-making among the spectators [99].

Another technology related to refereeing tasks is Goal Line Technology. This technology, first implemented in 2014, is primarily used to determine whether the ball has crossed the goal line. The system clearly indicates whether a goal has been scored, and if so, the referee is notified via their watch [25]. Like the semi-automated offside technology, Goal Line Technology, like the Semi-Automated VAR system, uses multiple high-speed fixed cameras under the stadium roof. However, the cameras used in this system are distinct from those used in the offside technology, as they are solely focused on each goal. Additionally, the inner inertia sensor, a vital component

of the offside technology, is not utilised in this case. This system also generates a 3D animation to be publicly displayed [27]. Among the prominent solutions in this domain are GoalRef [1], a goal-detection system developed by the renowned German research institute Fraunhofer; Hawk-Eye [2], a trusted partner of the Union of European Football Associations (UEFA) and Serie A (Italian Top League), acclaimed not only for its football applications but also for its collaboration with ATP and WTA in tennis; and GoalControl, or Vieww [3], which is a part of Sportec Solutions [4], a partnership collaboration with Deltatre and the DFL group, a specialist in sports and entertainment technology, and the *Deutsche Fußball Liga*(DFL), respectively.

3.1.2 Automated Broadcast Cameras

In competitions where live transmission is not economically profitable, for example, youth leagues or lower-tier divisions, it is less likely that some broadcaster is willing to produce and distribute a match. Furthermore, increasing competitions and contests often decreases broadcasting professionals' availability. However, having a transmission or record of these games might be helpful for fans who follow these teams or for coaches, analysts and scouts who can benefit from these game analyses.

A possible solution to these obstacles is using automated tracking cameras, providing a convenient alternative for broadcasting and recording these games by producing and streaming high-quality video without human intervention. Since this approach is a one-time investment, it would also be advantageous in terms of costs. Furthermore, these solutions can help address the issue of increased competition and matches, which often result in a need for more professionals. Usually, this software also offers additional benefits like coaching tools, analytics and match stats.

Pixellot [5] provides an AI-powered platform that automates capturing, streaming and analysing sporting events. It uses advanced Computer Vision and Deep Learning algorithms to track the players, the ball, and the action on the field [76] [46]. It can also adjust the camera's zoom, angle, and focus to create the best possible shots. The platform can reduce the cost and complexity of sports production, as it does not require human operators or multiple cameras. It can also enhance the viewing experience for the fans, providing dynamic and personalised perspectives of the game [77]. Moreover, it can improve the performance of the players and coaches, as it can provide video coaching solutions, video breakdown and analytics, and heatmaps [75]. The platform can work with different sports, such as football, basketball, ice hockey, baseball, softball, American football, rugby, volleyball, cricket, and handball.

Another platform that uses AI to automate the capture, stream and analysis of sports events is Spiideo [6]. It also tracks and adjusts the camera to follow the game's action and create optimal shots. It simplifies and lowers the cost of sports production, as it replaces human operators and

[1] https://www.iis.fraunhofer.de/en/ff/lv/net/projekte/goalref.html
[2] https://www.hawkeyeinnovations.com/
[3] https://vieww.com/about-us/
[4] https://www.sportec-solutions.de/en/index.html
[5] https://www.pixellot.tv/
[6] https://www.spiideo.com/

multiple cameras. It also improves the fan experience, offering dynamic and personalised game views [87]. Furthermore, it enhances the performance of players and coaches, as it provides video coaching, analytics, and heatmaps [86]. It also supports a diversity of indoor and outdoor sports like the previous solution [89].

3.1.3 Mediacoach and Beyond Stats

In an attempt to enhance performance, competitiveness, and attractiveness, as well as to democratise access to data and foster innovation and talent in the football industry, La Liga, Spain's premier football division, launched Mediacoach. The platform (Figure 3.2a) leverages artificial intelligence to offer real-time match data and video analysis to clubs, broadcasters, and fans [50].

Mediacoach has continually evolved for over a decade, adding new features and functionalities. One of its key features is the pre-match report, a comprehensive document detailing the opponent's statistics from the previous four matches. This report allows teams to anticipate game situations, analyse tactical positioning, and assess critical players [49]. It also provides exhaustive video analyses of past games, enabling teams to prepare effectively for upcoming matches. The platform delivers real-time performance data to the bench during matches through customised data dashboards [50] [49]. Additional features include a Tactical Camera and a post-match report with graphs, rankings, and statistics. Mediacoach supplies data and 3D graphics for broadcasters (Figure 3.2b) to enhance the match viewing experience [49].

In summary, Mediacoach is a dynamic tool that brings data-driven insights to the football industry's forefront, transforming how teams prepare, perform, and review their matches.

Building on the success of Mediacoach, LaLiga has taken a step further with the introduction of Beyond Stats. This innovative feature, developed in collaboration with Microsoft, uses metrics obtained from the Mediacoach platform supported by machine learning and artificial intelligence of Azure technology to process intricate player and ball-tracking data. Due to each La Liga stadium's optical and tactical cameras, Azure collects over three million data per match [51].

Beyond Stats offers fans and media access to 21 new metrics that delve into the technical and tactical nuances of the game [61]. These metrics include ball circulation, forward passes, line breaks, and scoring probability, providing a more comprehensive understanding of the match dynamics.

Like Mediacoach, Beyond Stats is designed to enhance the viewing experience of football. It offers a more granular analysis of player and team performance, which allows fans to gain a deeper appreciation of the strategies employed during the match and the skills exhibited by the players [52].

In essence, Beyond Stats is not just a data analysis platform; it is a tool that brings fans closer to the heart of the action, transforming the way they engage with the beautiful game of football. It is a testament to LaLiga's commitment to innovation and fan engagement, following in the footsteps of the successful Mediacoach platform.



(a) Mediacoach Platform (b) Beyond Stats broadcast Metrics

Figure 3.2: Mediacoach and Beyond Stats (Extracted from LaLiga, 2021 [49])

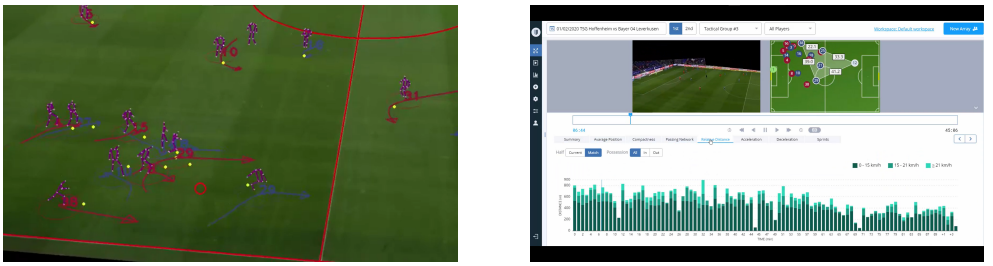
3.1.4 Track160

In the pursuit of democratising access to advanced sports analytics and fostering innovation in football, Track160 [7], an Israeli startup, has developed cutting-edge AI-powered software for comprehensive football analytics. Initially utilised in professional leagues, Track160 now focuses on amateur players and teams worldwide. The company's software can work with a single off-the-shelf camera and a cloud connection, enabling amateur football clubs to capture and analyse their data and improve their game [48].

This platform introduces two distinct solutions to revolutionise game analysis. *Coach160* [96] targets coaching staff, analysts, and team managers, automating detailed reports, tactical insights, and opponent analyses. Utilising a metaverse experience with a 3D game model facilitates match preparation and real-time decision support. On the other hand, *Player160* caters to individual players, focusing on personal performance metrics, skill enhancement, and training optimisation [97].

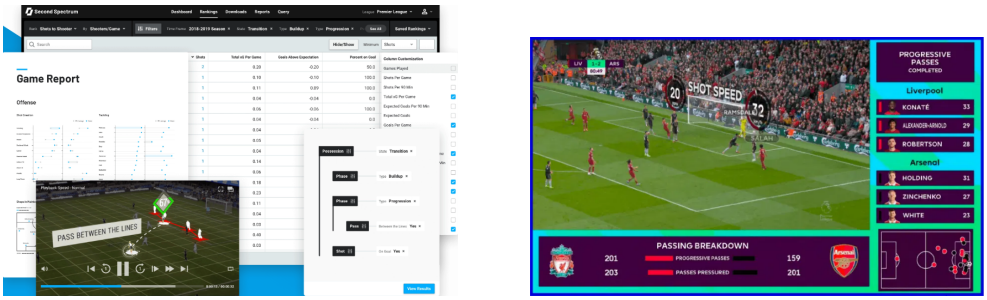
Track160 collaborates with some football and other tech companies, such as Microsoft and the *Deutsche Fußball Liga*, to leverage machine learning and large datasets for football analytics. The company also plans to expand and explore applications in other sports, aiming to make AI-driven sports analytics a universal tool for performance enhancement and player development [60].

[7] https://www.track160.com/



(a) Skeleton Tracking technology (Extracted from Track169, 2023 [97]) (b) Coach 160 Performance Analytics (Extracted from Track169, 2023 [96])

Figure 3.3: Track 160 Applications



(a) Second Spectrum Platform (Extracted from Second Spectrum, 2023 [85])

(b) Genius Sports Immersive Sports Data (Extracted from Genius Sports, 2023 [88])

Figure 3.4: Genius Sports and Second Spectrum

3.1.5 Genius Sports and Second Spectrum

Genius Sports [8] and Second Spectrum [9] are two entities that have made significant contributions to the sports industry, particularly in the realm of data analytics and visualisation.

Genius Sports is a global leader in digital sports content, technology, and integrity services. Their technology is used in over 150 countries worldwide, and they provide innovative and data-driven solutions to sports, media, and regulated betting markets. This includes providing cutting-edge software for collecting and distributing official data in real-time [88].

In May 2021, Second Spectrum, a fully integrated sports AI provider, was acquired by Genius Sports [79]. Second, Spectrum has chosen AWS as its preferred cloud and AI provider. The Second Spectrum platform (Figure 3.4a) captures real-time video and uses machine learning and Computer Vision technology to understand the events. This provides teams, leagues, and broadcasters with immediate and actionable insights. As the Official Optical Tracking Provider for the NBA and the Premier League, Second Spectrum generates accurate location data for several sports, enhancing understanding, evaluation, and content creation for teams and leagues [84].

The acquisition of Second Spectrum has allowed Genius Sports to enhance its existing products and services. Genius Sports' extensive network and operational scale, combined with Second Spectrum's innovative tracking and video augmentation products, has created a more decadent, more valuable official sports data offering [79]. This has driven fan engagement by providing a compelling experience that combines real-time data and analytics with innovative augmented video streaming and personalised content, as seen in Figure 3.4b.

One of the notable collaborations between Genius Sports and Second Spectrum is the development of the Premier League Data Zone. This feature, introduced by Peacock, a streaming service, incorporates live data analysis into broadcasts of select Premier League matches. It features live augmented graphics and overlays, visualising the data collected from live player tracking systems [64].

[8] https://geniussports.com/
[9] https://www.secondspectrum.com/index.html

3.1.6 Stats Perform

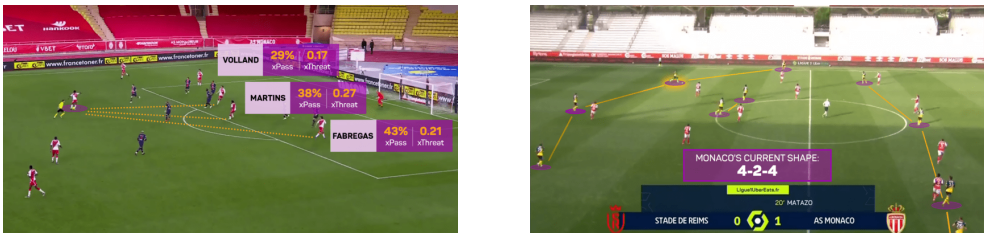
Stats Perform [10], trusted by industry leaders like SkySports and Google and by the Spanish and Italian League, play a pivotal role in enhancing sports analytics and fan engagement across various sporting disciplines.

In a similar vein to LaLiga's Mediacoach and Beyond Stats, Stats Perform, a global leader in sports analytics, has revolutionised the way sports data is analysed and utilised through its Opta Data. Opta Data provides advanced metrics that dig deep into the playing style of the entire team and the roles of their individuals, offering an unprecedented depth of analysis. This analysis can inform coaching strategies and assist player recruitment, transforming various facets of many sports [72]. Beyond data analysis, Stats Perform integrates cutting-edge technologies such as voice, VR, and AR, enhancing fan engagement and storytelling in sports.

Opta Vision, a significant innovation from Stats Perform, merges tracking data, high-quality event data, and robust AI to create a rich dataset. This dataset, built on a foundation of 40 years of extensive sports data collection, offers unparalleled insight into live play, delivering a new generation of football statistics [73] [70]. The platform provides a transformative approach to football analytics, integrating detailed event data. This event data includes passes and their probability of completion, as seen in Figure 3.5a, identification of line-breaking passes, shots along the respective expected goals and pressure intensity. This comprehensive view of the match also reveals team formations (Figure 3.5b), player movements, and tactical dynamics in real-time. This holistic analysis aids in tactical decision-making, player development, and scouting, offering invaluable insights for coaches, analysts, and scouts to understand and enhance team performance and strategy. Through Opta Vision, Stats Perform improves tactical decision-making and player development in sports, significantly transforming how the modern audience consumes and enjoys sports data.

Stats Perform covers the world of football in greater depth and detail and is used by some of the biggest partners in football, helping to inform and excite audiences globally with its comprehensive coverage of the sport. This AI-enriched data, available for over 40 men's and women's competitions, provides comprehensive insights, including on-the-ball pressure metrics, team shape, and passing data, crucial for recruitment profiling and player fitness metrics. The platform supports all stages of the analysis cycle, from pre-match opponent analysis to in-game performance monitoring and post-match evaluation. It also plays a critical role in player development across various age groups. Additionally, Stats Perform aids recruitment with an Opta-powered database covering 730,000 players across 250 competitions, supporting scouting operations worldwide. In the realm of player monitoring, Stats Perform offers next-generation wearable technology, enabling real-time tracking of player load during training and matches. This technology helps set customised alerts for player recovery and measuring physical performance, thus enhancing player efficiency and peak performance management. Their solutions empower clubs to manage scouting data and identify talents fitting specific recruitment profiles. Furthermore, their collaboration with

[10] https://www.statsperform.com/



(a) Opta Vision Pass Prediction (b) Opta Vision Team Shape Analysis

Figure 3.5: Opta Vision Use Cases (Extracted from Stats Perform, 2023 [71])

Electronic Arts (EA) highlights the broad applicability of Opta Data. EA has partnered with Stats Perform to optimise players' ratings, and PlayStyles features for the EA Sports FC 24[11] game, showcasing the use of Opta's bringing new levels of authenticity and realism to the game.

The Opta Stats Hub, for instance, provides an extensive repository of Opta stats for various competitions. This hub allows users to access the necessary data to assess team and player performance comprehensively, demonstrating its impact on understanding the game.

3.2 Emerging Trends in Football Video Analysis

In this segment, we examine the prevailing literature concerning Computer Vision applications in sports, particularly emphasising the challenges and methodologies of applying a single moving camera for football analysis. We first summarise the uses and areas of study for Computer Vision in sports. Then, we describe the mathematical principles and techniques for camera calibration, homography calculation, and coordinate mapping. We also compare and contrast the different approaches and evaluate their performance and accuracy.

3.2.1 Computer Vision in Sports

As we navigate through the digital age, the realm of sports is undergoing a significant transformation. This metamorphosis is primarily driven by advancements in technology, particularly Artificial Intelligence. The fusion of Computer Vision with sports is paving the way for a new era of automation and digitalisation, revolutionising how we perceive and engage with sports.

Thomas et al. (2017) [95] survey the state-of-the-art Computer Vision to assist with various aspects of sports in their article published in the journal *Computer Vision and Image Understanding*. Focusing on challenges like camera calibration, detection and tracking of players and ball and enhancing broadcasting with player modelling and motion analysis, it provides insights into the impact of using Vision in sports, especially in improving performance, officiating and broadcasting. Moreover, it overviews commercial systems used for those previous tasks. In addition, this survey provides a valuable resource for researchers and practitioners interested in the referred

[11] https://www.ea.com/games/ea-sports-fc?setLocale=en-us

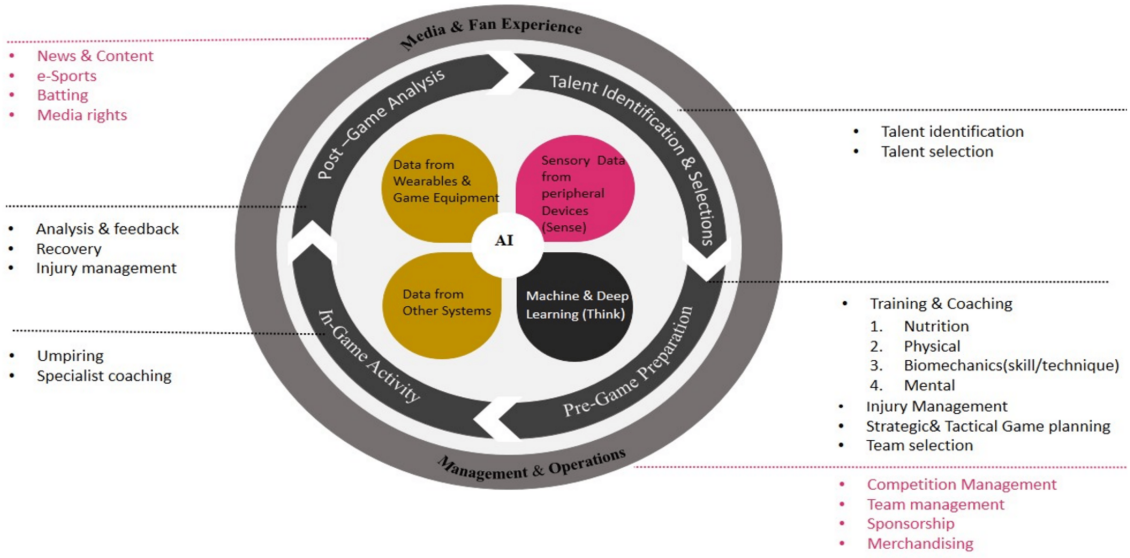


Figure 3.6: AI Role in Sports (Extracted from Naik et al. (2022) [63])

functionalities by listing publicly available datasets to develop and evaluate Vision tasks in various sports, such as football, basketball, tennis, cricket, and gymnastics.

A more complete and recent study by Naik, Hashmi, and Bokde (2023) [63] provides another comprehensive review of the use of Vision in sports. In this study, the authors delve into the potential applications of AI technology in the sports industry, such as chatbots, video highlights, analysis tools, and virtual assistants. They provide examples of how these applications can enhance the fan experience, improve player performance, and generate new revenue opportunities. To illustrate how AI can be used in sports, the paper presents a summary diagram of its usage, shown in Figure 3.6. The graph shows how a sports team uses intelligent systems for enhanced performance and operations. Data from games, equipment, and wearable devices is collected, covering player stats, biometric info, and videos. Algorithms process this data for patterns, predictions, and insights. The findings help in training, coaching, talent scouting, creating media content, personalizing fan experiences, and betting predictions. AI also boosts management efficiency, assists in umpiring, and provides specialized coaching feedback. This highlights AI and Deep Learning's extensive benefits in sports for players, coaches, fans, and management.

The authors review and evaluate different techniques for locating and identifying players in various sports, including colour-based, shape-based, motion-based, and Deep Learning methods and discuss various approaches for classifying and analysing the actions of players in sports videos using spatio-temporal features, recurrent neural networks, long short-term memory networks, and two-stream networks. Finally, the authors describe various methods for estimating and predicting the tactics of teams and players in sports videos. The methods include using Deep Extreme Learning machines, spatio-temporal data, visual analytics, and reinforcement learning. Figure 3.7 indeed presents a taxonomy of Deep Learning techniques, providing a structured overview of various methods used in the field. It categorises these techniques into four main applications: "Classification/Detection", "Unsupervised Detection", "Tracking", and "Trajectory Prediction".

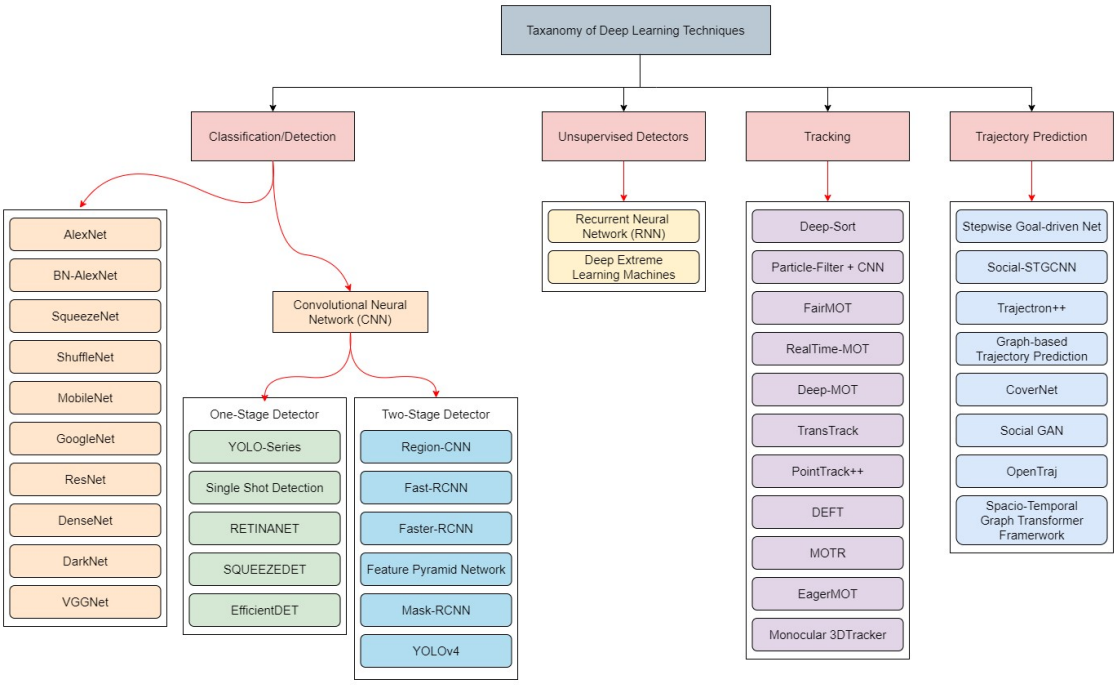


Figure 3.7: DL Techniques in Sports Applications (Adapted from Naik et al. (2022) [63])

These categories represent broad areas where Deep Learning techniques are commonly applied. Each category further encompasses specific methods, illustrating the diverse landscape of Deep Learning applications. This comprehensive and structured overview is a helpful reference for understanding the application of Deep Learning in various domains.

In conclusion, integrating Computer Vision and Deep Learning in sports is revolutionising the industry. These technologies are transforming how we collect, analyse, and use data in sports. As we continue to navigate the digital age, the fusion of AI with sports is expected to pave the way for a new era of automation and digitalisation, further transforming our perception and engagement with sports. Due to that, the potential applications of these technologies extend beyond the scope of improving performance, promising advances in areas like handling broadcast piracy [78] to deal with unauthorised sports streaming. These emerging applications further enhance the transformative power of Vision and Artificial Intelligence in this industry.

3.2.2 Calibration & Homography Calculation

One of the primary challenges of analysing football matches using Vision techniques lies in designing and implementing robust algorithms capable of calculating the homography or calculation matrices for the cameras. This is a critical step in mapping coordinates from the camera's perspective to real-world points. In other words, determine the geometrical relationship between the playfield and the image. In the following section, we will delve into these challenges, exploring current research methods and their effectiveness in tackling this problem.

3.2.2.1 Fixed or Moving Cameras

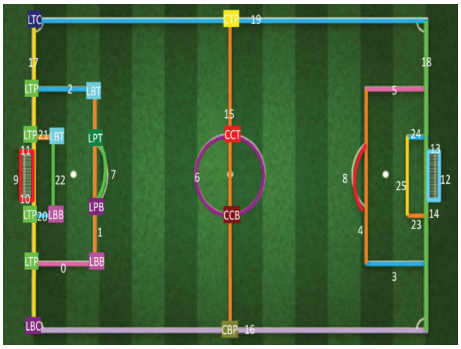
Before exploring methods and algorithms used for calibration, it is crucial to understand the unique advantages of different types of cameras. Manafifard [59] distinguishes different kinds of cameras in its camera calibration review for football videos.

Fixed cameras provide some advantages compared to others, like offering a stable and consistent view of the field. Due to their stationary state, the calibration process is generally more straightforward than moving cameras since focal length, sensor size, and position remain constant. Calibration techniques for this type of camera also allow pre-processing, such as using previously known reference points or consistent background modelling. However, it raises some challenges, like limited perspective, since this type of camera has a restricted field and can lead to potential gaps in data, which may also occur due to occlusions caused by crowded scenes. A standard solution to this problem is the use of multiple cameras that provide comprehensive field coverage due to the strategically placed cameras. This approach increases the field of view, improves resolution, reduces hidden regions, and simplifies calibration. However, it also raises hardware and computational costs, an issue we aim to address in our work. Moreover, an appropriate infrastructure is essential to effectively manage multiple cameras, which is not always feasible.

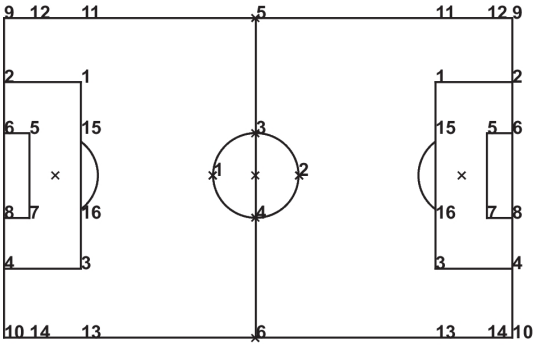
On the other hand, moving cameras, which are the primary focus of our investigation, offer solutions to some of the problems associated with fixed cameras. However, they also present their unique challenges. Moving cameras allow a dynamic coverage of football games, capturing close-up action and varying perspectives. This active coverage significantly enhances the viewer experience and offers a more engaging broadcast capture. Yet, the utilisation of moving cameras brings about a new level of complexity in calibration due to their continuous changes in position and orientation, posing a challenge in maintaining consistent calibration over time. Additionally, they can face occlusion issues and rapidly changing perspectives, which can complicate tracking tasks. As they capture different viewpoints, they encounter varying lighting conditions, further complicating the calibration process. Despite these challenges, various calibration techniques have been developed for moving cameras. For example, real-time camera tracking using sports pitch markings [94], offers high accuracy and real-time processing but is limited to specific environments and can face occlusion issues. Adaptive calibration and the use of auxiliary data such as gyroscopes or GPS to track the camera's activity provide flexibility and robust tracking in various environments, although they depend on sophisticated algorithms and high-quality sensors and can face errors or signal reliability issues [95].

3.2.2.2 Selection of Reference Points

One of the first steps in the calibration process, particularly for moving cameras, is the selection of reference points. They provide a known framework within the dynamic environment, aiding in mapping camera coordinates in real-world points. In sports, the field or game area offers a simple way to establish reference points for calibration. These points are known a priori across various sports where the playing area is well-defined and static and can be used effectively as references.



(a) Field Line Markers (Extracted from Karungaru et al., 2022 [43])



(b) Playfield Marks (Extracted from Manafifard, 2023 [59])

Figure 3.8: Football Field Reference Markers

For instance, in karate, where the competition mat is divided into coloured figures that serve as the playing field, the vertices of these figures are used as feature points for homography matrix calculation [91]. Similarly, squash utilises the playing field or court as a boundary, defined by a network of points connected by lines [14]. The court is visible in most broadcast and video recordings, so the intersection of those planes defines essential corner points and lines for calibration. Similar to football, basketball courts also provide a well-defined structure for calibration. In the context of basketball, Maglo et al. (2022) [57] have developed the KaliCalib framework, which is an innovative approach to sports-field registration. This method leverages key points sampled with perspective-aware constraints and includes the regression of basket positions to aid court localisation. This approach is particularly compelling given the varying colour designs and layouts of basketball courts, demonstrating the adaptability of the framework across different arenas.

Focusing on our main objective, football, some reference points are essential for an accurate homography matrix calculation. It consists of four field corners, four centre-line points, six points for the penalty box, four points for the goalkeeper area, and two points for the goalposts [43]. Discarding the centre line points, the others are duplicated since two symmetric halves form the field. These markings define the game boundaries and provide positional information beneficial for match analysis. The intersections of these lines, known as cross points, serve as excellent features for calibration, aiding in tasks like video analysis and player detection. The following two images, Figure 3.8, visually represent the football field. Comparing both images, the approach used in 3.8b is more complete using some extended penalty area points projected on the lateral line.

The process of extracting reference points or playfield involves several key steps, methodologies, and challenges, as detailed in various research papers. These steps are crucial in translating the physical characteristics of the football pitch into data that can be used for calibration.

Since the pitch markers are mostly lines, edge detection and line segmentation are vital in field extraction. Critical methodologies include the Hough Transform to estimate line orientation and position and the Probabilistic Hough Transform, which provides a more efficient line detection [59]. The Gridding Hough Transform (GHT) method further refines this process by dividing the

image into blocks and detecting lines within each block. Advanced techniques like RANSAC are beneficial for line detection amidst outliers, ensuring robustness against visual noise. In addition to line extraction, the extraction of arcs, such as the centre circle and penalty box arcs, present unique challenges due to their distinctive shapes and partial visibility. Arc extraction often employs ellipse detection techniques to distinguish different field regions and calibrate frames, especially in areas with few lines. Methods such as Hough-based techniques, RANSAC, and Least Squares Fitting (LSF) detect and fit the elliptical shapes of arcs [59]. Another method present by Berjón et al.(2023) [8] involves the use of stochastic watershed transformation for line mark segmentation. It treats an image as a topographical landscape, where pixel values represent heights. In this analogy, regions of an image are seen as drainage basins where rain would converge into the same lake. The technique simulates the flooding of this landscape. It starts at local minima (the lowest points, analogous to the bottom of basins) and gradually floods the landscape. As the water rises, it will eventually meet water from different minima, forming ridges or watershed lines at these points. The algorithm labels each unique landscape area where water from a single minimum would pool [11]. This approach is particularly effective over traditional edge detection stages as it efficiently filters out irrelevant edges caused by players, ball, or non-field elements. It is an advanced technique that segments regions in a grayscale image, interpreting the levels as terrain elevation. It lends itself naturally to flooding out irrelevant edges, enhancing the accuracy of line detection without the manual intervention of relevant regions.

Some more advanced techniques for field detection [93], particularly emphasising the role of semantic segmentation and using Convolutional Neural Networks (CNN). Semantic segmentation involves dividing the image into distinct regions based on specific criteria like object boundaries, which is vital for identifying field features. Another method of extracting pitch information is using color-based segmentation methods. Unless the analysis will be performed for games played on dirt fields, it is almost certain that the primary pitch colour is green due to the grass. Since the hue channel remains consistent across the area, providing a reliable feature for segmentation, Aléman-Flores et al. (2014) [2] emphasise using Hue, Saturation, Value (HSV) colour space instead of the traditional Red, Green, Blue (RGB). Then, an interval for the hue channel is selected, encompassing the typical colour values for the grass. Setting the correct thresholds makes it possible to isolate the playing field and classify the rest as background, like stands or advertisement panels. Then, the field area is expressed as a set of pixels whose hue value falls within that predefined interval.

3.2.2.3 Template Matching and Homography Calculation

Once the sports court is recognised, we must find the correspondences between the points and compute the homography matrix. In many sports applications, the Z-coordinate of the points on the playfield is considered zero [93] [2]. Therefore, the homography matrix simplifies the transformation.

Zhang & Izquierdo (2021 & 2023) [104] [105] have contributed with their work on camera calibration and homography estimation in sports videos, mainly focusing on the challenges

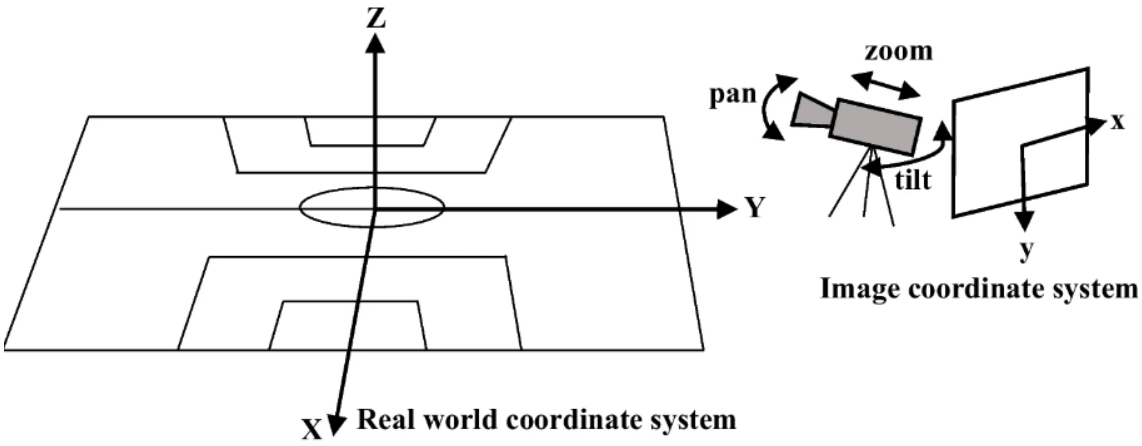


Figure 3.9: Pan, Tilt, Zoom Camera and Real World Coordinate System (Extracted from Manafifard, 2023 [59])

and methodologies inherent in this process. They dig into the process of homography estimation, which involves generating databases of various homographies and then searching for the best match for a given broadcast image. In the context of this work, these databases are collections of pre-computed homographies and features derived from sports field images with varying parameters like focal length, tilt angles, and pan angles built from This approach includes regressing the network to estimate the homography directly from the broadcast images. They then refine the homography estimation using spatial transformer networks to correct distortions to enhance accuracy and make them more precise.

An approach used for basketball and football proposes a novel method for calibration in broadcast videos [81]. They aim to find a homography, H, to register the target ground-plane surface from a broadcast video frame to a top-view field model. Since all three tasks can be accomplished with neural networks, their solution combines the three techniques integrated into a single network architecture. The first one, semantic segmentation, aims to extract key features of the field and remove irrelevant information for the homography computation, such as the audience, players, and other elements. A U-Net autoencoder with a cross-entropy loss function was employed for this task. It learns to classify each pixel in the broadcast image into one of the predefined regions of the football pitch. This network is trained using ground-truth labels generated by warping the top-down field model with the known homography. The second task was determining the camera pose initialisation process for calibrating broadcast videos, a sophisticated component of the overall methodology focusing on accurately determining the camera's position and orientation relative to the pitch. It assumes a pan, tilt, and zoom camera (PTZ camera) (Figure 3.9), ideal for its flexibility in capturing diverse views of sports events. This process generates a dynamic camera pose dictionary for the various possible camera positions. The dictionary is created through uniform sampling, which involves creating a grid of possible camera poses, or through a Gaussian Mixture Model-based clustering algorithm, which considers a broader range of camera configurations based on actual data patterns.

After establishing the initial camera pose parameters, they are refined using Zhang's' method and the Levenberg–Marquardt algorithm for more precise PTZ parameter determination. This refinement is crucial for accurately representing the camera's location and orientation. The final step in camera pose initialisation involves employing a Siamese network. This network assesses the input semantic image against each template in the camera pose dictionary, calculating the distance between the semantic segmentation of the broadcast video and each camera pose template to identify the closest match, thereby enabling precise alignment with the pitch model for enhanced broadcast quality. Homography refinement is the final stage and involves using a Spatial Transformer Network to adjust the homography matrix from the initial camera pose estimation, ensuring that the broadcast image is accurately matched to the field model and enhancing the accuracy of their method.

Tsurusaki et al. [98] presents a sophisticated homography estimation and evaluation approach, allowing highly accurate camera calibration. The process also begins with generating intersection pairs between the input image and the field template using point-matching methods. For each set of intersection pairs, a homography matrix is calculated. This matrix projects points from the pitch real-world coordinates to the image coordinates. The evaluation function seeks to maximise the overlap between the projected field lines, using the AND operator on the result of the estimated homography and the actual field lines in the input image while minimising the non-overlapping areas using the exclusive or (XOR) operation. After several estimates, the one with the better evaluation is selected as the final result.

This sports field registration method [56] represents a significant advancement in computer vision, particularly for sports analytics. Combining a novel network architecture with perspective-aware keypoint sampling, self-supervised learning, and advanced data augmentation, this method aims to provide more accurate and robust homography estimations even in challenging scenarios like fast camera movements or limited visibility of pitch features. This approach is precious in sports broadcasting and analysis, where accurate pitch mapping and player tracking are crucial.

Yao et al. [103] presented another approach to calibrating a moving camera in the context of football. The proposed method involves three main steps. The first step involves detecting and tracking distinctive feature points in each frame of a football video. Harris corner points are identified as feature points in each frame. They are tracked across frames using the Kanade–Lucas–Tomasi (KLT) tracker, creating a trajectory matrix, $A_{2\mathrm{KN}}$ represents the motion paths of these detected feature points, N, across K frame. A sub-matrix, $D_{2\mathrm{TM}}$, is extracted from the previous one using a sliding window approach. This sub-matrix focuses on feature points that are consistently visible, represented by M, over T frames. The trajectory matrix is then analysed and decomposed into two matrices: one representing camera motion (low-rank matrix L) and the other individual motion (sparse matrix S). This decomposition, formulated as an optimisation task, assumes that the camera motion will be a dominant, coherent pattern. In contrast, individual motions, like independent players, will be less frequent and sparse. Using the decomposed trajectory matrix L, which predominantly represents camera motion, a homography matrix is estimated for each pair of consecutive frames. This is done using the Direct Linear Transformation

(DLT) method based on pairs of corresponding feature points extracted from *L*. The homography matrices allow for the calibration of each frame concerning the field ground.

Another research, even though it is applied to American football, presents two methods for estimating and tracking homographies in a moving camera scenario [67]. The first one uses player and helmet tracking, provided by RFID tags used by players. It starts with identifying the players' helmets based on motion, shape, colour or other features that enable helmet distinction. After this, the system checks which helmet tracks persist over the subsequent three frames, which is crucial to ensure that data tracking is reliable and that players are consistently present. The system then extracts the corresponding player bounding boxes in that last frame for helmets that have been consistently tracked for three frames. These bounding boxes are used to identify specific points representing player positions on the ground plane of the video frame. Usually, the mid-point of the base of each bounding box represents where the player is standing. Using the correspondence between the midpoints on the ground plane in the video frame and the player points in a 2D plane, the system calculates the homography. The system uses a cost metric for the innovative sampling-based homography computation method to evaluate the computed homography matrices. The matrix best aligning the pre-registered data points with the corresponding points in the video frames is selected. This system has limitations, like motion blur or fast camera movements, which can hamper helmet tracking performance. The computation only works if more than four helmet tracks persist across the three subsequent frames.

The second method uses key point extraction on the ground plane using SIFT and then FLANN to match the keypoints between subsequent frames. The inverse of the homography matrix of the initial frame transforms all detected keypoints in the current video frame to corresponding points in the 2D plane coordinate system, aligning them with the real-world positioning. The correspondence between these transformed keypoints and their positions in the following video frame's ground plane is used to compute the homography for the subsequent frame. The homography matrix captures the transformation needed to map points from the 2D plane to the new video frame. The method's success depends on having a sufficient number of keypoints.

In 2024, Sommers et al. presented a novel task in sports analytics that goes beyond merely tracking and identifying athletes on the field. More than simply tracking and identifying the game's intervenients, it also displays the game data in a minimap. This representation, also called game radar, offers a 2D perspective of the ongoing game state. The process involves calibrating the camera using TVCalib [93]. This calibration is a two-step process: first, the pitch is localized using semantic segmentation, and then the camera parameters are estimated by minimizing the reprojection errors of pitch segments iteratively. Once the camera is calibrated, its corresponding homography maps player positions onto the 2D minimap. This mapping determines the relationship between the camera's view and the actual field, allowing for the identification of the visible area within the frame [83].

In an effort to address the limitations of previous methods, a novel calibration pipeline was proposed. This pipeline enables camera calibration using a 3D football field model. It extends the process to assess the multiple-view nature of broadcast videos. The approach begins with a key-

point generation pipeline derived from the SoccerNet dataset annotations, taking advantage of the geometric properties of the football field. Following this, classical camera calibration is executed through the Direct Linear Transform (DLT) algorithm without further refinement, thus relying on the robustness of the geometric properties for precise calibration. This algorithm is a classical camera calibration method that calculates the projection matrix from 3D object points to their 2D picture projections. The method has been extensively tested on real-world football broadcast datasets such as SoccerNet-Calibration, WorldCup 2014, and TS-WorldCup. The results demonstrate superior performance in multiple and single-view 3D camera calibration while maintaining competitive results in homography estimation compared to state-of-the-art techniques [37].

A significant challenge of most of these methodologies is insufficient points within the frame to compute the homography matrix. This issue arises because its accuracy heavily depends on identifying and matching keypoints between the 3D world image and the reference pitch model. This limitation can lead to inaccuracies in the homography calculation, thereby affecting the precision of subsequent camera calibration and alignment processes essential for sports analytics and broadcasting applications.

3.2.2.4 Optical Flow in Sports's Context

In scenarios where traditional calibration methods struggle, mainly when key features like corners or edges are obscured or insufficient, optical flow emerges as a viable solution to estimate camera motion. This is often the case in dynamic environments like football matches. For instance, essential features such as the penalty box or the centre circle might not always be visible during a game. Relying solely on these for calibration, even if there are four identifiable points to calculate a homography, can be problematic. This is because, in a fast-paced game, these points might not provide a consistent reference, leading to inaccurate or misleading results [54]. In contrast, optical flow can be used to map coordinates, offering a more robust approach using movement patterns. It is possible to infer how the camera moves by tracking the movement of pixels or features between frames. This is particularly useful in dynamic scenes like football videos recorded by a moving camera [59]. There are many techniques to estimate motion using optical flow, like correlation or block matching, feature tracking, energy-based or gradient-based approaches, using pixel intensity, and polynomial expansion transform-based methods [45].

In this context, using RANSAC plays a pivotal role in enhancing the accuracy and reliability of homography estimation. In the homography estimation process, many outliers may be found. These can be due to moving objects in the scene or mismatches in the correspondences identified by the optical flow algorithm. RANSAC is critical in this scenario as it is specifically designed to cope with a significant proportion of outliers [36]. Moreover, the presence of outliers is even more pronounced when using PTZ cameras, which involve complex movements like panning, tilting, and zooming. Using RANSAC in these circumstances allows for a reliable homography estimation despite those challenges. In addition, an iterative application of this robust approach contributes significantly to the overall accuracy of the flow estimation, as evidenced in the RANSAC-FLow research of Shen et al. [82], which evaluates the performance on various datasets and its ability

to handle dense and sparse correspondences. It further enhances the application of RANSAC in particularly complex scenes where a single homography is insufficient, like two widely different views, which is common in the context of football game analysis using multiple cameras placed in other points of view or two video frames, which is more suitable for our moving camera analysis.

A commonly used method to estimate optical flow is the Lucas-Kanade solution [42]. This approach estimates optical flow locally around each pixel by assuming that the movement of pixels is essentially constant in a small neighbourhood. This assumption holds exceptionally well when the frames are captured at a high frame rate and is particularly useful for players' features in images that will be analysed later. It solves the flow equations of the pixels in a neighbourhood using a least-squares approach, which tries to find flow vectors that best fit the actual motion in a least-squares sense. This method works by minimising the accumulation sum of the squares of the difference between the observed and predicted values. The main advantages are the speed and efficiency, which makes it suitable for real-time video, and the accuracy in time derivatives since it is good at capturing temporal changes. On the other hand, it can struggle at the boundaries of moving objects where the assumptions of constant flow within a neighbourhood may not hold. In terms of dense optical flow, one of the state-of-the-art methods is Recurrent All-Pairs Field Transforms (RAFT). RAFT is a deep network architecture that extracts per-pixel features, builds multi-scale 4D correlation volumes for all pairs of pixels, and iteratively updates a flow field through a recurrent unit that performs lookups on the correlation volumes. It has shown strong performance [92].

A recent study on players' movement analysis presents a semi-automatic approach to estimating homographies [22]. In their research, Działowski and Forczmański use a panning camera. It is similar to the previously shown PTZ camera, but this moves exclusively horizontally. Their technique begins with pitch segmentation by defining a colour range in the HSV colour scheme. This is crucial as it isolates the colour of the pitch grass by creating an interval of greens. However, the existence of outliers, for example, a green object in the stands, or when the pitch colour is outside the expected range, can raise a problem. This issue is addressed using Generative Adversarial Networks (GAN), which helps define the playing area accurately. Then, objects like the game intervenients and the ball are eliminated using the closing operation, which consists of a dilatation followed by erosion, effectively filtering out these objects and enhancing the focus on the playing area. The camera angle is then obtained using dense optical flow. At the same time, optical flow estimates the vector fields that pinpoint the velocity of each moving pixel in time-varying image sequences, and dense optical flow records each pixel's motion in consecutive frames [45]. This calculation indicates the camera movement since the camera panning is the significant cause for pixel positioning changes between different frames. This optical flow can be defined from the equation:

$$I(x,y,t) - I(x+\Delta x, y+\Delta y, t+\Delta t) = 0, \tag{3.1}$$

where the intensity of a pixel at a specific position (x, y) and time instant (t) is denoted by (I).

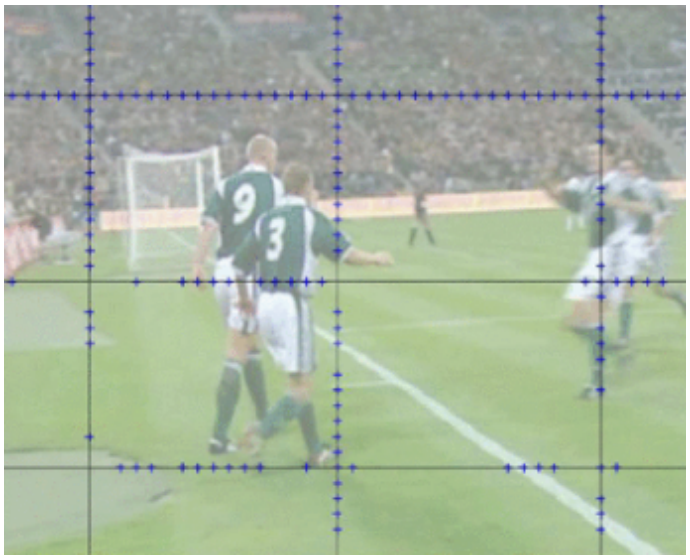


Figure 3.10: The Six Straight Lines used as Features (Extracted from Felip et al., 2008 [24])

The changes in position are represented by Δx and Δy, showing the movement in the x and y coordinates, respectively. In contrast, Δt represents the time increment between the frames [20]. A mask of the pitch is applied to the frames so that the optical flow calculations are focused on the game area rather than on irrelevant information like the sky or the audience. The mean optical flow vector is calculated for every frame, representing the camera movement for a certain frame. The accumulation sum of these vectors over time enables the approximation of the camera angle relative to its initial position. We can select these frames by analysing the frames where the camera angle significantly changes compared to the initial orientation. These key frames identified using the optical flow are then used for camera calibration. This involves selecting corresponding points between the camera image and the field model, allowing one plane to be mapped to another.

3.2.2.5 Encoded Motion in the Realm of Sports

In analysing dynamic sports environments like football matches, the MPEG codec presents a unique approach for motion estimation, distinct from the previous optical flow analysis techniques.

Felip et al. [24] proposed an approach to use MPEG Data, precisely the motion vector field from the compressed stream. In other words, using the fixed grid MacroBlock (MB) structure in MPEG frames to create lines for homography estimation, MPEG video compression involves dividing frames into blocks or MacroBlocks. Each MB can have an associated motion vector, which indicates how the pixels in that block move between frames.

In video frame alignment, a feature typically refers to a distinct and recognisable part of the image that can be easily tracked between frames. These features should be invariant to the transformations expected between the frames and robust to noise. The feature selection process considers only the MBs with an associated motion vector. This choice is crucial because these motion vectors provide essential information about how each part of the frame moves. The method involves

selecting six straight lines as features. The lines are chosen based on three rows and columns of MBs, as shown in Figure 3.10. The selection is manual, emphasising picking lines representing the frame's structure and motion, thus facilitating accurate alignment. The manual selection process allows human judgment in choosing the most representative lines. However, it also introduces a subjective element, which could impact the reproducibility and consistency of the method across different frames. The primary aim of selecting these features is to facilitate the estimation of a homography that aligns one frame to another. In this case, these features, represented by lines, serve as reference points or guidelines for the algorithm to calculate the transformation needed to align the frames.

The feature correspondence process follows the previous step. This step is crucial for maintaining the continuity and consistency of features across frames, which is essential for accurate frame alignment and homography estimation. It is about translating the initial features from the reference frame to their new positions in the target frame, guided by the motion vectors. This involves adding the motion vectors associated with the MBs to the points defining the initial features in the reference frame. The success of this step hinges on the quality and reliability of the motion vector data extracted from the MPEG stream. The algorithm predicts where those features would be in the target frame by adding the motion vectors to the initial feature points. This step estimates the movement or transformation from the reference to the target frame. It is critical for accurately aligning the frames, as it directly impacts the following homography estimation.

As already mentioned, video processing, especially tasks involving motion and compressed streams, can be noisy with a high percentage of outliers. The Helmholtz Tradeoff Estimator (HTE) is a robust statistical method designed to handle situations with significant outliers in the data. It is advantageous in scenarios where traditional estimation methods, like least squares mentioned in an approach using optical flow [42], might fail due to the high noise level. HTE tries to find a compromise between the variance of the inlier set and the number of samples in this set. This balance is crucial because focusing solely on minimising variance can lead to overfitting outliers, while solely maximising the number of samples might include too many outliers. HTE starts by taking multiple random subsets of the data. Each subset classifies data points as inliers or outliers based on specific criteria. It evaluates each model by considering the variance and the number of inliers. The best model is the one that strikes the optimal balance, as defined by the estimator's criteria.

An equation system is solved using the least median squares method to find the homography matrix. This step involves applying the homography theorem to corresponding features (lines) identified in both frames and solving for the homography matrix using a robust statistical method like least median squares. This method minimises the median of the squares of the differences between actual and predicted values, making it less sensitive to outliers than traditional least squares. The equations derive from the line correspondences between the two frames. Each line correspondence provides a set of constraints formulated into equations.

3.3 Relevant Datasets

Data collection is a crucial step in developing a video analysis system for football matches, as it provides the necessary input for the algorithms that will be used to train and test their performance. To achieve this step, we also need to research the existing relevant datasets created and used by previous studies in this field. There are several existing datasets for football video analysis. We will review some of them and discuss their limitations and challenges for our specific scenario.

3.3.1 IPL Azadi Soccer Dataset

The Azadi Soccer Dataset [13] [31] [32], hosted by the Intelligent Processing Lab (IPL) from the Sharif University of Technology, is a specialised dataset for research in the field of computer vision. The dataset is primarily intended for academic research and development.

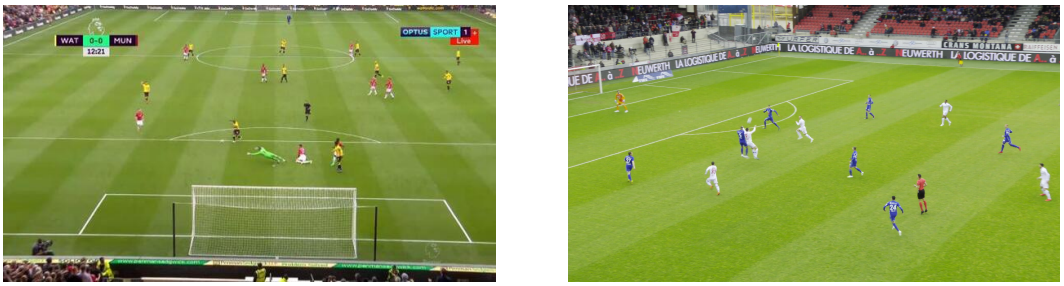
This dataset is tailored explicitly towards applications in football video, such as player tracking. The high-resolution footage allows for precise monitoring of players' cameras throughout the game, making it invaluable for detailed analysis. Event recognition is another critical application of this dataset, as it can identify and analyse key events like goals, fouls, or offsides. Finally, the dataset is instrumental in developing algorithms for calibrating cameras in sports settings. This calibration is essential for accurately mapping the 2D video footage onto a 3D pitch model, a process crucial for advanced analyses, such as tactical evaluation and game strategy development.

These cameras are strategically placed to capture various angles and aspects of the matches, providing comprehensive coverage. The dataset also has a detailed football field model, illustrating the cameras' precise locations and orientation. This complete camera setup not only aids in diverse research applications but also presents unique data processing and analysis challenges. Particularly notable is the inclusion of the moving camera, which offers excellent opportunities for more challenging calibration research. This aspect of the dataset aligns perfectly with our research focus on developing robust calibration techniques for dynamic camera movements.

3.3.2 SoccerNet Dataset

SoccerNet [34] [19] is a large-scale dataset designed explicitly for analysing football broadcast videos. It is primarily used in the computer vision field, particularly for sports analytics. It is used for research purposes and algorithm development.

The data includes 550 broadcast videos in the Matroska Video (mkv) format at 25 frames per second (fps) with two different resolutions, 720p and 224p. It offers many video clips from numerous matches, often captured from broadcasted contests. This includes wide-angle views of the field and close-up shots of players and specific events, as well as a wide range of scenarios, team formations, player actions, stadium environments, and lighting conditions. SoccerNet provides annotations for various game events, like cards, substitutions and goals, with a temporal marker indicating the time in the video where the specific occurrence occurred, which helps evaluate models in event detection and activity recognition. In addition to the basic annotations, SoccerNet-v3



(a) View of a camera behind the goal

(b) Broadcast frame captured by the main camera

Figure 3.11: SoccerNet Dataset Data

introduces a comprehensive set of spatial annotations. These include detailed line and goal annotations, bounding boxes for players and objects, and multi-view player correspondences [18].

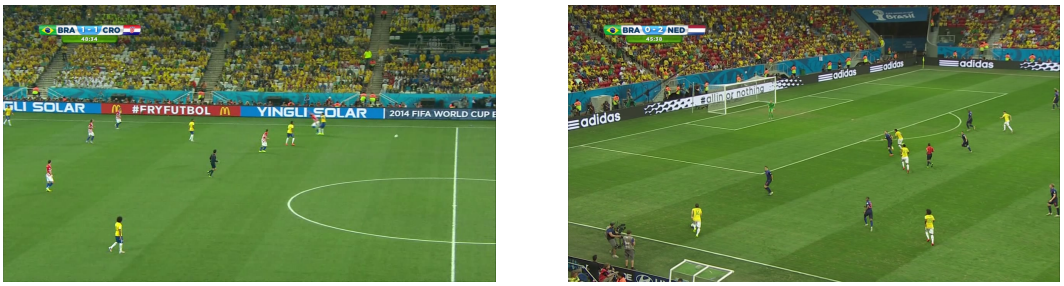
The dataset has been instrumental in various tasks divided into three main themes: Broadcast Understanding, which includes action spotting, replay grounding, ball action spotting, dense video captioning, and camera shot segmentation; Field Understanding, focusing on field localisation and camera calibration; and Player Understanding, with tasks like tracking, re-identification, and jersey number recognition. Our primary focus aligns with the last two themes, emphasising the importance of camera calibration to identify and track players and the ball in sports analytics. The following image, Figure 3.11 presents two examples of the available data provided in this dataset.

3.3.3 World Cup 2014 Dataset

The 2014 World Cup Dataset is a specialised collection of images derived from football games held in Brazil during the 2014 World Cup. This dataset has been specifically curated to advance research in computer vision, mainly focusing on tasks such as field detection and grass segmentation. It was assembled by Homayounfar et al. [38] in their work, which consisted of localising a sports field from a single image of the game broadcast.

This dataset includes a total of 395 images from 20 different games of the 2014 World Cup. Each image is annotated with ground truth data, which comprises homographies (the transformation of the image's perspective) and grass segmentations (identification of grass areas in the image). These images were captured in nine unique stadiums under diverse conditions, including day and night, and sometimes under rain and heavy shadows, offering a broad spectrum for analysis. This ground truth data is crucial for tasks like calibrating the perspective of field images or segmenting different areas in sports analytics. This data set is divided into a training and validation set, which includes 209 images and a test set of 186 images. This division is essential for testing the accuracy of the models trained on this dataset. A notable feature for expanding the dataset is the possibility of flipping images horizontally, effectively doubling its size [105], a technique often used in computer vision to enhance dataset variability.

One of the main disadvantages of the World Cup 2014 dataset is that the images are not temporally ordered and are single, static images like those shown in Figure 3.12. This means that it is not



(a) View of a camera behind the goal (b) Broadcast frame captured by the main camera

Figure 3.12: WordlCup 2014 Dataset Examples

suitable for video and sequential analysis but for static image analysis. However, the vast number of pictures and detailed ground truth labels make it an invaluable resource for testing calibration algorithms and checking their accuracy against diverse conditions.

3.3.4 TS-WorldCup Dataset

The TS-WorldCup Dataset significantly advances football-related datasets for computer vision and machine learning applications [17]. Developed as an extension of the original WorldCup dataset, the TS-WorldCup dataset stands out due to its substantial size and time-sequence information.

This dataset comprises 3812 field images meticulously extracted from 43 videos of the FIFA World Cup tournaments held in 2014 and 2018. The dataset's size is noteworthy, ten times larger than its predecessor, the WorldCup dataset. Each image in this collection features detailed field markings, providing a rich data source for field detection and analysis.

A distinctive aspect of the TS-WorldCup dataset is its incorporation of time-sequence information, which sets it apart from static image datasets. This temporal aspect allows for more dynamic and comprehensive analysis, particularly suited for tasks involving motion tracking, event sequencing, and temporal pattern recognition in football games. In addition to the features as mentioned earlier, the TS-WorldCup dataset introduces a unique element of spatial differentiation since it distinguishes the images on the left (Figure 3.13a) or right (Figure 3.13b) halves of the field. Furthermore, the dataset includes homography ground truth data.

The dataset is structured into two primary subsets: 2925 images designated for training purposes and 887 images for testing. This division facilitates the development and evaluation of machine learning models, ensuring a balanced approach to training and validation.

3.3.5 SportsMOT Dataset

The SportsMOT dataset consists of 240 high-definition video clips collected from over 100 match tapes from three popular sports: basketball (Figure 3.14b), football (Figure 3.14a) and volleyball [101]. These videos were collected from competitions like the Olympic Games or the NBA on YouTube. Each video, with a resolution of 720P and a frame rate of 25 FPS, is carefully trimmed to an average of 485 frames, ensuring no shot changes within a clip.



(a) Example of a left annotated video frame



(b) Example of a right annotated video frame

Figure 3.13: TS-WorldCup Data

The dataset includes outdoor football games and indoor basketball and volleyball matches to capture various sports environments. The perspectives in these clips are diverse, ranging from the standard sideline views in NBA games to unique serve zone angles in volleyball and aerial views in football. This variety in settings aims to bolster algorithmic adaptability across different sports tracking scenarios.

SportsMOT encompasses 240 clips, each averaging 495 frames (approximately 19.8 seconds). The dataset is divided into training, validation, and test sets containing 45, 45, and 150 videos, respectively, with no game duplication across splits.



(a) Football frame example



(b) Basketball frame example

Figure 3.14: SportsMOT Data

Chapter 4

Research Proposal

Within this chapter, we will outline our approach for computing homography in football matches using moving cameras. We will discuss our main ideas and our planned solution to achieve the objectives mentioned earlier.

4.1 Problem Definition

Accurate homography calculation plays a pivotal role in various applications in sports, particularly in dynamic and complex environments such as football matches. With this research, we aim to overcome several challenges that current calibration methods face, particularly when key features such as corners or edges are insufficient to achieve an accurate result.

It is worth noting that most of the existing methods for calculating homography in static football images do not effectively consider temporal consistency when applied to video [41] [37]. These methods face significant challenges and often become computationally intensive as they are applied frame by frame to individual video frames instead of considering the temporal consistency [65].

Furthermore, the effectiveness of these methods relies on the visibility of a minimum of four keypoints [37]. However, in areas of the field with few keypoints, such as those shown in Figure 4.1a, accurate homography calculation can be problematic. Even if four identifiable points are present, precise homography computation is not guaranteed, particularly in challenging lighting conditions like shadows [15], as shown in Figure 4.1b. Moreover, players, the referee, or other interferences could obstruct these reference points. These methods also heavily depend on accurate segmentation and may struggle with inaccuracies due to the lack of robust data points if not regularized to handle outliers [93]. Consequently, in areas with sparse keypoints, homography calculations are more likely to be inaccurate, ultimately affecting the precision of the transformation.

We will explore motion-based approaches to overcome the challenges associated with sparse keypoints in video-based homography calculation. By leveraging camera motion, we aim to enhance the reliability of homography estimations in dynamic football match scenarios, seeking to address those limitations.



(a) Example frame with few pitch marks (SoccerNet Dataset[34] [19])



(b) Some keypoints are occluded by light conditions (TS-Worldcup Dataset[17])

Figure 4.1: Challenges faced for homography calculation in football

4.2 Proposed Solution

The main focus of this work is to analyse camera motion by using different inputs to understand and map movement within a given video sequence. We will explore two main approaches.

Our proposed approach involves initially calculating the homography using manual methods. This manual homography will serve as a reference point for comparison, creating a benchmark against which we can compare the state-of-the-art techniques and our proposed approaches. It will also be used as the reference first frame homography in the proposed methods. This procedure ensures a high level of confidence in the accuracy of the point correspondences, as it allows one to choose what is perceived to be the ground truth directly. Moreover, by manually mapping an input image to a football field template, we will gain a better understanding of how these transformations work between the two surfaces. This hands-on experience will help us comprehend the intricacies of the transformation and provide a baseline for validation.

The first approach involves creating a superimage panorama using the camera motion. This computational technique transforms individual images into a cohesive whole, ensuring each video frame is precisely mapped onto a common plane. The basis for this computation is the homography between the first frame and the pitch plane. The goal is to predict the homographies of subsequent frames using only this initial transformation matrix, supplemented by motion data, which will also used to create the super image. This first approach hypothesises that by knowing the homography of the first frame and incorporating the calculated motion, we could generate a comprehensive view of the video. This method would potentially not only provide a complete panorama but also enable precise mapping of each frame within the sequence. The superimage would then serve as a reference for recalculating homography, ensuring continuity and consistency across frames. As illustrated in Figure 4.2, given the initial frame and its corresponding homography, by applying the camera motion detected between the two frames used to construct the panorama, we should derive the homography of the second frame, thereby facilitating the seamless incorporation of successive frames into the superimage.

The second approach integrates motion data directly into the homography equations. Given that we know the keypoints in the first frame, we will test if we can apply the calculated motion to these keypoints to determine their new positions in subsequent frames. By recalculating the

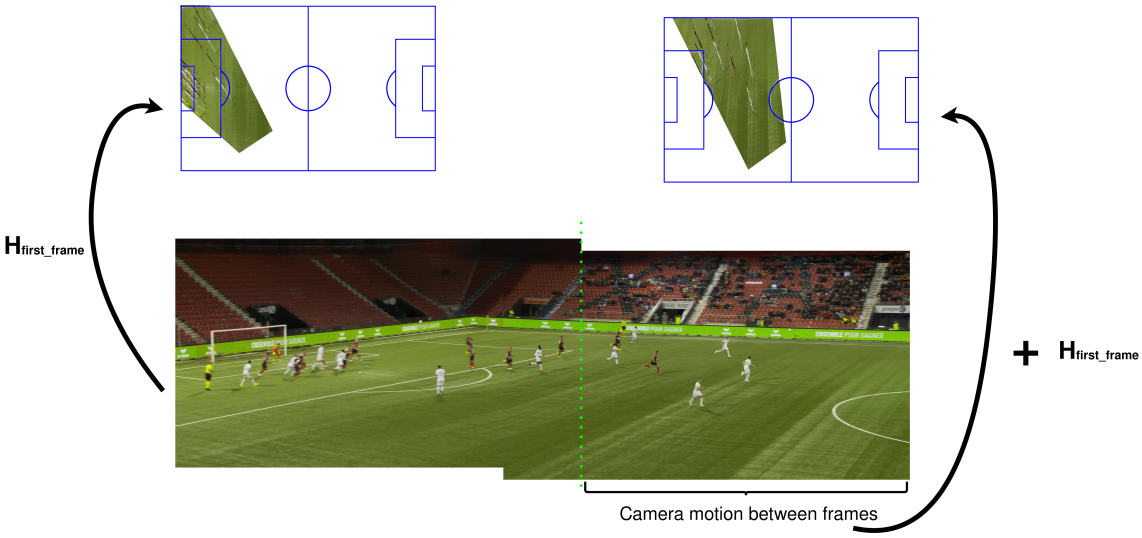


Figure 4.2: Proposed approach using the panorama

homography using these updated positions, we will experiment to see if it is possible to track specific points on the pitch accurately. Since the destination points (the points on the pitch template) remain constant, this approach could potentially allow for ongoing and precise adjustments of the homography transformation.

To calculate motion, we will employ different inputs to understand and map movement within the video sequence. Specifically, we will use optical flow, a technique often used in sports for tracking applications, such as player or ball tracking. Optical flow involves calculating detailed motion information between consecutive frames at the pixel level, allowing us to track the movement of pixels. Additionally, we will employ motion vectors derived from MPEG-4 videos, which provide a block-based representation of motion and can be used to estimate overall movement in the scene. We aim to compare these pre-calculated motion vectors with optical flow data to see if MPEG motion vectors can be as effective while saving time and reducing computational costs. Lastly, we will manually annotate points and their respective distances to serve as a benchmark for validation and comparison against the automated approaches. The motion information obtained from these inputs will be used to calculate homography, allowing us to effectively align and analyse the video frames.

Chapter 5

Methodology

This chapter explains the methodology and experimental setup employed in our study.

First, we will describe the strategy used for developing the proposed methods. This involves creating the super-image and manipulating the homography equations using various techniques to obtain motion information. We will also outline how the homography transformation is obtained and tested. Each method is described in detail, from the initial concept through to execution. We will also discuss the rationale behind selecting each technology and technique for a specific method, respective implementation details, and how it integrates with other processes within the study.

Next, we will explain the evaluation metrics chosen to provide the context for understanding how the results are measured and compared. We will discuss each metric in detail, highlighting why these particular metrics were selected and how they contribute to assessing the performance and accuracy of our methods.

Finally, we provide a complete description of the data and hardware utilised for this proposed approach. This includes the original data sources, the ground truth data available, and the ground truth data we created.

5.1 Implementation Strategy

This section describes our implementation strategy for evaluating motion-based homography methods. Our approach encompasses several critical phases.

First, we analysed and integrated state-of-the-art methods to benchmark against our approaches, ensuring a thorough comparison across different scenarios. Next, we implemented manual homography calculation, establishing a baseline for accuracy and validation.

For our proposed methods, we incorporated motion data derived from multiple sources: manual input, optical flow analysis, and MPEG-derived motion vectors, enabling further analysis on the different inputs.

Building on these foundations, we implement super image techniques and manipulate equations. This involves leveraging the calculated motion to recalculate the homography transformation continuously, ensuring alignment with the evolving dynamics of sports video footage.

5.1.1 Automatic State-of-the-Art Methods

This section, provides a detailed explanation of the two state-of-the-art models that were used for comparison. Using different methods allows for comparison with diverse approaches and mitigates bias that could arise from evaluating only one method.

5.1.1.1 TVCalib

TVCalib [1] presented a unique approach compared to other state-of-the-art implementations. It proposes a novel 3D sports field registration method, which treats it as a camera calibration problem instead of the usual sports field registration as a homography estimation problem. The used framework for registering the 3D model of the sports field can be divided in two main parts: segment localisation that performs instance segmentation and selects appropriate points with respective labels from a known calibration object like a 3D model; the calibration module, which is the main contribution of this work, predicts camera parameters by iteratively minimising the segment reprojection loss [93].

The source code is available in a GitHub repository, and we have utilised their readily available inference code in our solution. This code is crucial for identifying points on the field and calibrating the camera, allowing for accurate 3D registration of the sports field. The inference model used initialises a DeepLabV3 model with a ResNet101 backbone for semantic segmentation. It loads a pre-trained model from a checkpoint and performs inference on an image batch. We have adapted the original inference code to suit our needs better and extract the most valuable information for analysis.

5.1.1.2 Sports Field Release

Just as we did with TVCalib, we also employed the inference method provided by Jiang et al. [41] in their research. They proposed a framework that leverages optimisation techniques for registering sports fields from broadcast videos, a divergence from the commonly used feed-forward paradigm.

The framework can be divided into two stages. The first estimates an initial transformation matrix using a typical feed-forward paradigm performed by a deep neural network. The second stage uses this initial estimate to warp the sports field template to the current frame. The warped image is concatenated with the frame, and the registration error is calculated through a second neural network. The registration error represents the difference between the actual and predicted positions of points in the sports field template when mapped onto the frame. This error is then

[1] https://mm4spa.github.io/tvcalib/

backpropagated to the homography parameters to obtain the gradient, which indicates the direction in which the parameters should be updated to minimise the registration error. In essence, the system adjusts the parameters until the predicted positions of points in the template align as closely as possible with their actual positions in the broadcast video. This optimisation-based inference allows the method to outperform a typical feed-forward setup in terms of accuracy.

We reused the available code to calculate the homography matrix and visualise the optimisation results.

5.1.2 Manual Homography Calculation

As already referred, we used a manual approach to serve as a reference point for homography computation. This will work as a benchmark for comparing the state-of-the-art techniques and our proposed methods.

Initially, an image is selected to be read and used as input along with the predefined template image of a football field. Then, reference points are manually selected for both the input and template images through a graphical interface. In this step, the number and order of the chosen points in both images must be precisely the same. These keypoints are stored and used to compute the homography matrix. This transformation aligns the input image with the template. We employed the least squares method for the homography calculation. This method minimises the squared differences between the observed keypoints in the input image and their corresponding transformed positions on the template. By solving a system of linear equations derived from these correspondences, the least squares approach iteratively adjusts the parameters of the homography matrix to optimise alignment. The primary reasons for this choice were the small set of manually selected points since there are few football field markings, and this number is even smaller in a football video frame since it only captures a small part of the field. Additionally, the points used for calculating the homography are manually selected. This manual selection process ensures a high level of confidence in the accuracy of the point correspondences, as it allows choose what is perceived to be the ground truth directly. This controlled selection process inherently minimises the risk of outliers, making the least squares method a suitable and effective approach for this context.

5.1.3 Motion Calculation

Before exploring the specifics of each method, we will provide a brief explanation of some foundational topics that are common to all methods. This will help establish a solid understanding of the basic principles underlying our approach to camera motion calculation.

Camera motion can be briefly described as the displacement of a specific point from one frame to another in a video sequence. This displacement captures the change in position of any given object, person, or place in the field of view. Although, and since we are analysing such dynamic scenes as the football matches are, it is important to distinguish between camera motion and object movement within the video sequence. The displacement from the camera might be mismatched

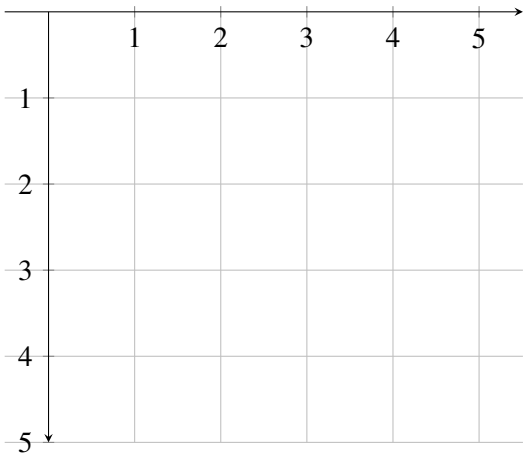


Figure 5.1: Pixel coordinate system

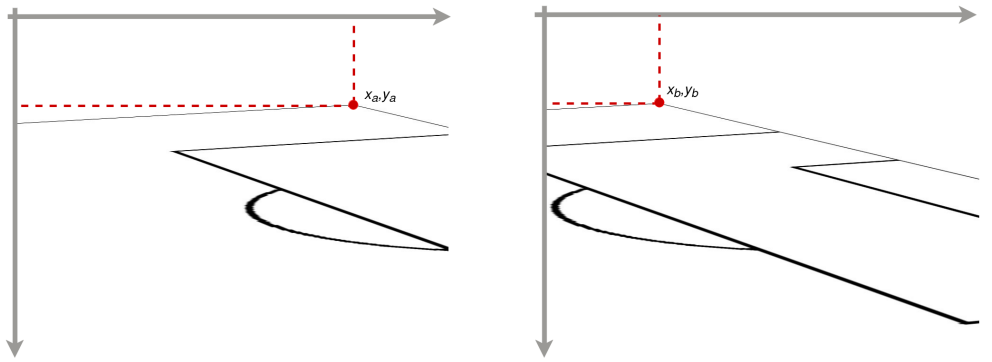
or confused with the movement of objects, but these are separate phenomena. Camera motion refers to the change in position of the camera's point of view, while object movement refers to the actual motion of objects within the frame. To comprehend camera motion without misleading these concepts, it is important to analyse static entities between consecutive frames. Tracking this movement over time offers valuable insights into the game's dynamics. It helps understand the perspective from which the game is viewed, the areas of the field that are focused on, and the overall flow of the game as captured by the camera.

The proposed approach uses the mean displacement of all pixels to average the movement across the entire image, which is crucial when inferring camera motion. Using the mean displacement allowed us to concentrate the motion on the camera and minimise noise introduced by players, the ball, and other elements moving in different directions. It generates a coherent displacement pattern across all pixels, serving as a reliable indicator of camera motion. Additionally, due to perspective, objects situated further from the camera require fewer pixels to represent the same real-world distance travelled than objects closer to the lens. As they appear smaller in distance their movement occupies less space on the sensor. Considering the mean displacement, we account for these perspective shifts and obtain a more accurate measure because it inherently accounts for depth variation within a scene. It provides a holistic view of movement across the frame, smoothing anomalies caused by individual object movements or depth-related distortions.

In image processing, the coordinate system differs from the traditional Cartesian coordinate system usually used in mathematics. In an image, the origin (0,0) is located at the top left corner, meaning that the x-coordinate (horizontal axis) increases as we move right. In contrast, the y-coordinate (vertical axis) increases as we move down , as represented in Figure 5.1.

With this understanding, we can validate the following:

- **Positive X displacement** — This means that the x-coordinate of a specific point in the second frame is lower than the x-coordinate of the same point in the first frame. This indicates a left shift, meaning the camera moved to the left.



(a) Scenario A - Corner area coordinates: (x_a,y_a) (b) Scenario B - Corner area coordinates: (x_b,y_b)

Figure 5.2: Relation between X-axis displacement and camera motion

- **Negative X displacement** — This indicates that a particular point's x-coordinate in the second frame is greater than that point's x-coordinate in the first frame. This represents a camera movement to the right.
- **Positive Y displacement** — This shows that the y-coordinate of a certain point in the second frame is greater than the y-coordinate of that same point in the first frame. Since the y-axis increases downwards, the camera moves up.
- **Negative Y displacement** — This last scenario indicates a decrease in the y-coordinate value between consecutive frames, denoting a downward movement of the camera.

To understand the previous statements , let us analyse Figure 5.2. In Figure 5.2a (Scenario A), we can see the identified corner area as (x_a, y_a). Likewise, Figure 5.2b (Scenario B) represents the same point labeled as (x_b, y_b). It is clear that $x_b < x_a$. If Figure 5.2a is followed by Figure 5.2b, it becomes evident that this represents a shift of the camera to the right, as the goal area becomes more visible. The displacement is calculated as $x_b - x_a$, which results in a negative value since $x_b < x_a$. Conversely, if the displacement was positive, it would mean a leftward camera movement between Scenario B and Scenario A.

5.1.3.1 Motion Calculation from Manual Input

The manual process for selecting points closely mirrors the approach used for homography computation. The first step is to choose points on both images, ensuring the selection order is consistent across both frames. This consistency ensures that a point determined in the first image corresponds directly to one selected in the second image.

When selecting the points, it is necessary to note that calculating the camera motion requires selecting static points. Therefore, objects that might shift from one frame to another, such as the ball or players, are inappropriate. For achieving better results in the calculated motion, the most effective points for selection are field marks or the stadium's structural elements. Field marks are advantageous as they are easy to identify and remain static throughout. Structural elements could

include features like an entrance to the stands, the substitutes' bench, or advertisers that typically surround the field. However, it is essential to note that these advertisers should not be electronic, as they are often animated and may introduce inaccuracies in the motion calculation.

After selecting the points, they are processed to calculate the motion. A dictionary is constructed to facilitate easier debugging and analysis. This dictionary records the coordinates on the first and second images for each point. Additionally, it stores the displacement, which encompasses the shift along the x-axis, the y-axis, and the total displacement. The total displacement is computed using the Euclidean distance formula, providing a comprehensive measure of movement between the two images. Since the number of points assigned is typically low in this context, using a dictionary proves to be beneficial despite the processing involved. Although the complete displacement information was helpful in providing insights and analysis on the calculations as the first approach, it was later found that for the optical flow and encoded motion vectors, only the essential information was used as these approaches involved a more significant number of pixels.

$$Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5.1}$$

where the (x_1, y_1) represents the coordinates of the first point and (x_2, y_2) the coordinates of the second point. After collecting information for each pair of correspondences, we also computed the mean x displacement, the mean y displacement and the mean total displacement, which were the values calculated from the Euclidean formula.

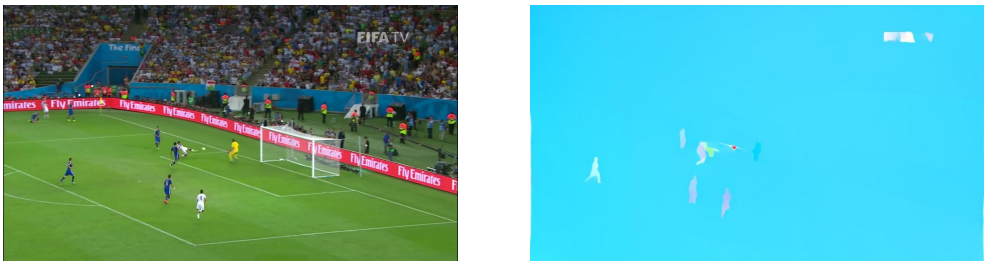
5.1.3.2 Motion Calculation from Optical Flow

To explain the approach for calculating camera motion using optical flow, it is essential to understand the methodology employed for optical flow computation.

In this approach, we utilised the RAFT model, a state-of-the-art technique for estimating optical flow [92]. The objective was to compute the flow between each pair of consecutive frames in a video. To achieve this, we adapted an example provided by PyTorch [2].

An optical flow model commonly uses two images as input, with the predicted flow as the output. By mapping each pixel in the first image to its corresponding pixel in the second, this flow shows how each pixel has shifted. We initiated the process by reading the video and converting it into a format compatible with the model. Subsequently, we applied the PyTorch implementation of the RAFT model, which sticks to the architecture outlined in the original paper [92]. The RAFT model generates lists of expected flows, where each entry, N, represents a batch of predicted flows (N, 2, H, W) associated with a specific model iteration. Our interest lies in the final predicted flows, which are the most accurate in this context, so we focused on the last item in this list. This left us with a tensor of dimensions (2, H, W) where each H represents the y-coordinate of the pixel

[2] https://pytorch.org/vision/0.12/auto_examples/plot_optical_flow.html#sphx-glr-download-auto-examples-plot-optical-flow-py



(a) Original Frame (b) Visualising predicted flows

Figure 5.3: Comparison of the original frame and the visualisation of the optical flow

and W represents the x-coordinate. The tensor's first element represents the pixel's horizontal displacement, while the second element indicates the vertical displacement. This effectively captures the displacement of each pixel in both the horizontal and vertical directions from the first image to the second. We applied the model to each pair of frames in the video, generating a series of predicted flows. These predicted flows were then stored for subsequent use in calculating camera motion. In Figure 5.3 we can see an example of the visual representation of the predited flows obtained by this model and the respective original frame.

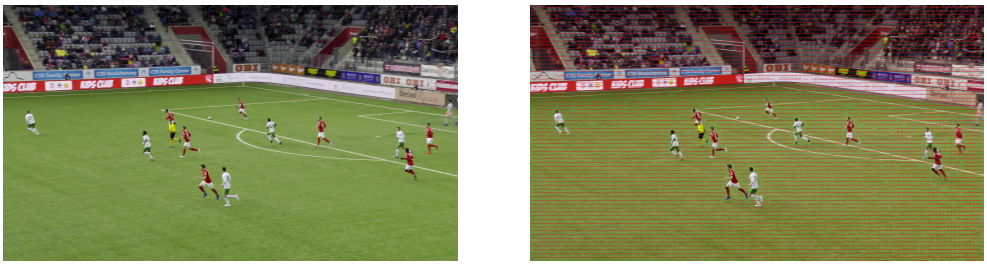
In the specific context of football matches, the displacement between consecutive frames is often minimal. Therefore, we can compute the total displacement between two non-consecutive frames. In the case of manual selection, we can directly select and calculate the motion. On the other hand, when using optical flow, the stored flows are between consecutive frames. Therefore, if we aim to calculate the total displacement between, say, the first and the tenth frames, we cannot simply select these two frames and compute the motion. Instead, we must accumulate the motion using all intervening frames. This involves summing up the displacements from the first to the second frame, the second to the third frame, and so forth, up to the ninth to the tenth frame.

To calculate the motion, we iterated between any given two frames and computed the displacement for the current pair of frames as the mean of both horizontal and vertical displacements in the flow. The total displacement was then updated accordingly. Ultimately, we obtained a pair of values representing the mean horizontal and vertical displacements.

A notable divergence from the manual approach was that we retrieved only the mean displacement instead of using a dictionary. This was justified in this particular case, as we had more points than those selected using the manual approach, which would lead to slow and inefficient computation if we were to create a similar dictionary.

5.1.3.3 Motion Calculation from Encoded Motion Vectors

For the approach where we used motion vectors from MPEG-encoded videos, we used the Motion Vector Extractor provided by Bommes et al. [12]. Motion vectors, used in video encoding like MPEG, represent the direction and magnitude of apparent motion in a video scene by matching macroblocks between frames. They are crucial for reducing data size, which is why this approach may represent an innovative strategy to reduce the complexity of the optical flow.



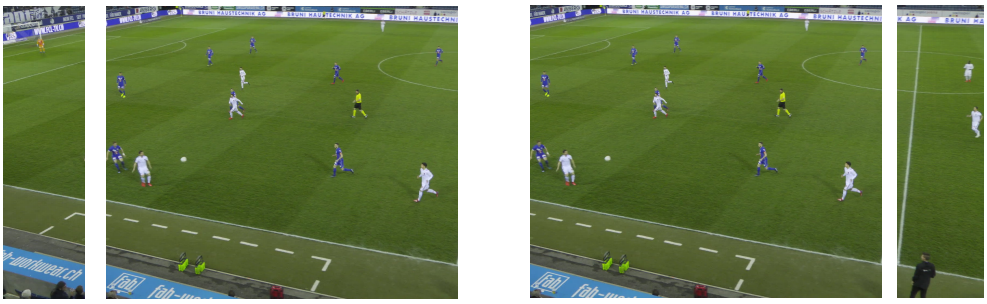
(a) Original Frame (b) Motion Vectors

Figure 5.4: Comparison of the original frame and the visualisation of the motion vectors extracted from MPEG encoding

A single command is executed with the video as the input to extract the motion vectors. The output is a list of files that encapsulates essential information about each frame. It includes the reference frame, source points, and destination points. These points outline the initial and terminal locations of the apparent motion in the video, offering valuable insights into the dynamics within the video frames. We then processed the retrieved information and the motion vectors for each frame in a video, storing them in a structured format (a JSON file). This allows easier access and analysis of the motion vectors later on since the array contains more information than what we need. Figure 5.4 provides the visual representation of the motion vectors obtained in this process.

The displacement calculated by these motion vectors can be calculated in two ways. The first method computes the displacement by measuring the difference between the macroblock's location in the destination frame (the current frame being analysed) and its original location in the source frame (the reference frame from which it has moved). This approach generates a direct pixel displacement value in integers. The second takes into account the motion vector's quantisations and scaling factors applied during video compression. For each x and y component of the motion vector, the calculated motion is divided by a scaling factor known as the motion scale. This process translates the quantised motion vector components, which are typically stored as integer values to save space, back into a precise floating-point representation that can reflect sub-pixel movements. This added precision often makes this method more accurate, thereby enabling a more refined analysis of movement within video frames, which is particularly useful when analysing videos that require high fidelity in motion representation.

The motion calculation step in this approach strongly resembles the optical flow method, with the key difference being that this encoded motion is already calculated instead of the optical flow, which might bring advantages in reducing computation complexity and efficiency. Suppose our goal is to calculate the total displacement between two non-consecutive frames. In that case, we also need to adopt a cumulative approach, as we only have information about the source and destination point coordinates between consecutive frames. We iterate between any given two frames and compute the displacement for the current pair of frames as the mean of both horizontal and vertical displacements in the flow. The critical distinction here lies in the type of data available. While the optical flow provides the displacement between consecutive frames, in this case, we



(a) Camera moving left, from B to A (b) Camera moving right, from C to D

Figure 5.5: Concatenation Example

need to calculate the horizontal and vertical displacement for each pixel coordinate. The output, once again, is a pair representing the displacement in the x and y axes.

5.1.4 Super Image Creation

Upon obtaining the motion information, either manually, from the motion vectors or the optical flow, we started this approach by outlining the operational mechanism of the panorama. Initially, we determined the information we wished to retain or discard from one frame to the next. The basic solution was to concatenate any information that still needed to be provided to the current panorama. In other words, if a new frame has a part of the field that has not been seen yet, we add it to the panorama.

Recalling the previous relationships between positive and negative displacements on both the x and y axes, we incorporated these into our methodology. For the panning movement or displacement in the x-axis, the approach involves concatenating the new portion of the second image onto the left or right of the first frame, depending on the camera movement direction. So, the new panorama size would be the current panorama width plus the displacement in the x-axis. This displacement corresponds to the part of the new frame containing the latest information, which will be cropped. Figure 5.5 shows an example of how it works. In the first case (5.5a), when the camera moves left, *B* represents the current panorama state or the initial frame, and *A* represents the cropped portion of the new frame that has not appeared in the video yet. The total horizontal displacement equals the width of image *A*. The final panorama is created by stitching these two images together. In the second example (5.5b), we can see a similar situation but for a rightwards movement. The new information, *D*, also shows a part of the field not yet seen in the video, and it is stitched to the right of *C*, which represents the current state of the panorama. Again, the horizontal displacement equals the width of the cropped image, in this case, *D*.

The handling of tilting, or motion along the y-axis, presents a unique set of challenges. It is important to note that this type of motion is less noticeable in football videos, as the camera's positioning and field of view often minimise these movements. Indeed, these movements do exist, and addressing the y-axis was more about aligning the frames than introducing new information,

as the horizontal shifts were almost always more significant than the vertical movements. In cases of noticeable movement, we continued to crop the new information in the horizontal axis. Then, we adjusted the new segment with the current panorama based on the vertical shift.

We initiated by comparing the motion between a single pair of frames to validate and implement our approach. We began the analysis with non-consecutive, based on the observation that camera displacement between consecutive frames is typically minimal, making the study of such frames a complex work.

For instance, one of the challenges we encountered in this first approach was the need to align images of varying dimensions optimally. To address this, we used a technique that involved padding both images with black pixels. The direction of the vertical movement difference determined where these pixels should be placed. Pixels were added to opposite sides of the pictures depending on whether this difference was positive or negative. This method proved effective in compensating for discrepancies in image shape, thereby facilitating the concatenation process.

This focused approach served as a valuable starting point for our subsequent steps but also offered crucial insights for debugging and refining our methods and preparing for more complex tasks, in this case, analysing a complete video.

The following step involved applying these techniques to a video sequence after validating the concatenation mechanism for both the right and left sides.

Compared to the single-frame approach, the primary distinction was the need to manage scenarios where the camera movement did not yield any new information or when the camera's direction changed. This included cases where the camera remained static between the two analysed frames or moved in opposite directions. Still, the movement in one direction was insufficient to introduce new information compared to the initial frame. To address these scenarios, we employed a tracker for the x-axis to record the displacement relative to the first frame. This ensured that the origin of the panorama corresponded to the origin of the first frame, allowing us to effectively manage cases of static or insufficient camera movement. Additionally, by tracking the current x value, we could determine when it was necessary to update the panorama, verifying if it falls within the expected range tracked so far. If the current displacement lies outside this range, it indicates a significant movement that might require alignment. The total range of displacements was then updated to include the new current position, helping to track the maximum and minimum horizontal displacements encountered so far. This approach helped maintain alignment and avoid redundancy in the final panorama.

The second main difference was dealing with the previously mentioned camera's movements in opposite directions. To handle these situations, the tracker stores a direction flag based on the horizontal displacement. The direction is set to $left$ or $right$, accordingly. If a shift in direction is detected, meaning that the sign of the displacement does not align with the stored flag, then two critical updates are necessary. The first and most apparent is that the direction flag must be updated to reflect the new movement direction. This ensures that future displacements are correctly interpreted in the context of the new direction. The other is to adjust the current x value by considering both the calculated displacement and the frame width. When changing direction,

including the frame width ensures that the tracker accurately accounts for the complete shift from one side of the frame to the other. This is necessary because the displacement must reflect the new position beyond the current frame's bounds. This prevents errors that could arise from ignoring the frame's dimensions during a direction change.

This methodology can offer several advantages for homography calculation. Firstly, creating a panorama from successive frames provides a broader and more comprehensive view of the field. This enhanced field coverage is particularly beneficial for tactical analysis and player positioning, offering a holistic perspective that a single-frame analysis might miss. Secondly, constructing a panorama reduces the dependence on computationally intensive keypoint detection and matching across multiple frames. Instead, we calculate the homography for the initial frame and use the camera motion data from subsequent frames to update this transformation. This approach streamlines the process, making it more efficient and effective for homography calculation in dynamic scenarios such as a football match.

5.1.4.1 Homography Computation Using Super Image Motion Displacement

In line with our exploration of the panorama creation and image alignment processes, we scrutinise the complexities of homography calculation. This computational technique transforms individual images into a cohesive whole, ensuring each frame is precisely mapped onto a common plane.

The basis for this computation was the homography between the first frame and the pitch plane. Our goal was to use this initial transformation matrix supplemented by motion data to get the homographies of subsequent frames.

We needed to apply the motion information to the homography to achieve this. We utilised homogeneous transformation matrices for this purpose.

$$\begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \tag{5.2}$$

where Δx and Δy represent the displacements along the respective axes. By incorporating these displacements, we created a translation matrix that represents the movement.

To integrate this motion into the homography, we combined the translation matrix with the initial homography matrix. This combination is not straightforward; it involves matrix inversions to correctly apply the translation in the context of the existing transformation. Specifically, we invert the initial homography matrix before applying the translation, then invert the result again to obtain the final new homography matrix. This ensures that the transformations are correctly compounded.

In practice, we used a tracker to guide the calculations and construct the panorama. This tracker stores information about the panorama width according to the first frame. For any information, stitching to the left will have negative horizontal coordinates, while stitching to the right

will have coordinates greater than the frame's width. Additionally, the tracker keeps the current x value, allowing us to verify if the information is repeated. The tracker also includes a direction flag indicating whether the video moves to the left or right. This ensures that the correct adjustments are made when the direction of movement changes, maintaining accurate alignment across frames. Using this was the way of ensuring that the horizontal position stays within the bounds of a single frame's width. This is crucial for maintaining consistent and accurate displacement calculations, especially when dealing with large movements across multiple frames.

5.1.5 Directly Applying the Motion Information Into the Homography Equations

In this second approach for calculating the homography of video frames, we also leverage the first frame and motion information. However, instead of using the information used to create the panorama, we directly applied the motion displacements to the homography equations.

This equations were provided using the Direct Linear Transform (DLT) algorithm. This method is used to find the homography matrix that relates two sets of points in different planes, as explained in 2.1. It starts by forming a matrix A that represents the transformation. This matrix is obtained by

$$\begin{bmatrix} -x_i, & -y_i, & -1, & 0, & 0, & 0, & u_i x_i, & u_i y_i, & u_i \\ 0, & 0, & 0, & -x_i, & -y_i, & -1, & v_i x_i, & v_i y_i, & v_i \end{bmatrix} \tag{5.3}$$

where each pair of corresponding points in the source frame are represented by x_i,y_i and by u_i,v_i in the destination frame. This generates the A matrix, which has a shape of $2nx9$ where n is the number of pairs of correspondences.

To determine the homography matrix from A, we employed Singular Value Decomposition (SVD). SVD is a mathematical technique that factorises a matrix into three distinct matrices: U, S, and V^T. Here, U and V are orthogonal matrices, and S is a diagonal matrix containing singular values representing the magnitude of each corresponding eigenvector.

The homography that transforms source points into destination points is represented by the last column of the matrix V. This column corresponds to the smallest singular value in S, which is associated with the solution that minimises the total squared reconstruction error.

In our specific case, we already have the homography matrix for the first frame. This homography was calculated using a set of keypoints (corresponding points between the frame and a reference template). The initial set of keypoints used to calculate the first frame's homography serves as our reference for calculating the homography for subsequent frames. These keypoints consist of source points from the first frame and corresponding destination points in the template. For each subsequent frame, we calculate the displacement using either MPEG motion vectors or optical flow data. The source points from the first frame are updated by adding the computed displacement, simulating their new positions in the current frame. The destination points, which

are the correspondences in the field plane, remain the same. This is because the template does not change, and we aim to align each frame to this fixed reference.

After setting all this up, the methodology followed a similar process to our initial approach. We iterated through the given video sequence, calculating the motion between non-consecutive frames using either MPEG motion vectors or optical flow data. This motion information was then used to update the source points for each frame. By applying the DLT algorithm, we recalculated the homography matrices for each frame. By consistently updating the source points based on motion information and using the fixed destination points, we can compute the homography for each frame. This approach ensures that each frame is aligned to the coordinate system of the first frame, facilitating motion analysis and frame alignment along the video sequence.

5.2 Metrics for Evaluating Homography Estimation

We have chosen specific evaluation metrics to measure and compare the results of our experiments. In this section, we will discuss each metric in detail, explaining why these particular metrics were selected and how they contribute to evaluating the performance our methods. Understanding these metrics is crucial for interpreting the results presented in later chapters.

5.2.1 Intersection-over-Union

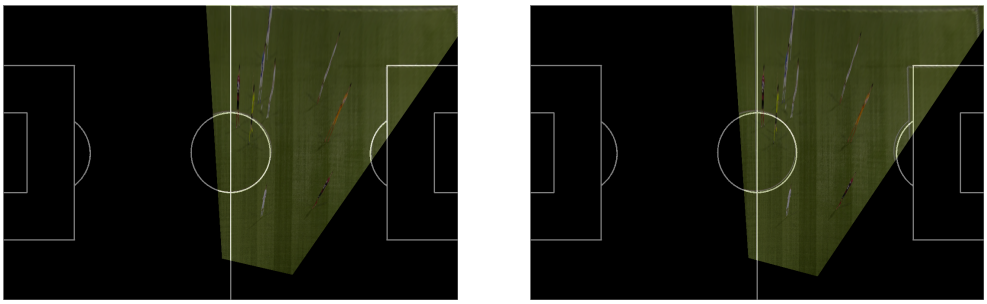
The first metric we applied in our study is the Intersection-over-Union (IoU). This metric is a standard measure in object detection tasks, where it quantifies the overlap between two bounding boxes - the predicted bounding box and the ground truth.

However, the IoU is not limited to object detection tasks; it is also applicable in segmentation tasks, which is the context of our study. In segmentation tasks, instead of bounding boxes, we work with masks.

The IoU is calculated as follows: First, we compute the union and the intersection of the two masks. Then, we divide the area of intersection by the area of union to obtain the IoU score (refer to Equation 5.4). This score provides a quantitative measure of how well the predicted mask aligns with the ground truth mask.

$$IoU = \frac{(\text{Ground-truth Mask} \cap \text{Predicted Mask})}{(\text{Ground-truth Mask} \cup \text{Predicted Mask})} \tag{5.4}$$

In our specific scenario, we compute the Intersection over Union (IoU) results as follows. We apply the predicted and the created ground truth homography transformations to a given frame. This process creates a projection of the frame onto the pitch template, as illustrated in Figure 5.6. Using this projection, we generate binary masks for both the predicted and ground truth regions. Finally, we calculate the IoU by comparing these two binary masks.



(a) Frame Projection Using Ground-Truth (b) Frame Projection Using Prediction

Figure 5.6: Annotated points example and respective generated mask

5.2.2 Distance between players' positions

We also used the Mean Squared Error (MSE) as a critical metric in our work. The MSE calculates the average squared error, which is the sum of the squared differences between our predictions and the ground truth, divided by the sample size.

Our dataset provides ground truth for the players' positions both in the video frame and on the 2D pitch plane. We took advantage of this information for evaluation purposes. Specifically, we computed the MSE of the distances between the ground truth 2D positions and the predicted 2D positions. These predicted positions were obtained by applying the estimated homography to the players' positions in the video frame.

Given the Euclidean distance between the corresponding pixels in the 2D plane, we convert it to meters using a scaling factor. Given that the template size was 1050x680, we divided the distance by 10 to match FIFA's recommended pitch size of 105x68 meters. The MSE was then calculated by summing these squared errors and dividing them by the number of players in a specific frame. The following equation can represent it:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(P_i^{gt} - P_i^{pred})^2 \tag{5.5}$$

where N represents the total number of players in a frame, the error, calculated by the Euclidean distance, is represented by $|P_i^{gt} - P_i^{pred}|$, where P_i^{gt} shows the ground truth 2D position and P_i^{pred} shows the predicted 2D of the i^{th} player.

This metric offers a more detailed evaluation of how closely the predicted homography aligns with the true homography. By applying the predicted homography to the player positions and comparing the result to the ground truth positions, we can quantify the accuracy of the homography estimation. In other words, a good homography estimation should result in transformed player positions that closely match the ground truth positions.

We chose the MSE over other metrics, such as the Mean Absolute Error (MAE), because it provides a more detailed measure of the prediction errors, emphasizing the impact of more

significant errors. It is also more appropriate to detect variability, which was particularly useful for understanding how the predictions work across the video duration.

5.3 Data and Hardware Overview

In developing our video analysis system for homography calculation in the context of football matches recorded by moving cameras, we have explored a variety of potential datasets. These datasets, each with unique features and applications, could provide valuable input for our research.

In our data selection process, we considered several factors. One of the key considerations was the accessibility and availability of the data. While some datasets were freely available, obtaining them required contacting the author or dataset creator or signing specific agreements to gain usage rights. Given our time constraints, we prioritised datasets that were immediately accessible rather than waiting for approvals or permissions. Another critical criterion in our data selection process was the data format. Specifically, we required video data rather than static images. This requirement was due to our research and development objectives involving dynamic analysis that could only be achieved through video data. However, in our work, we ended up using both video format and individual frames extracted from these videos. Nevertheless, it's important to note that static standalone images, which are different from using video frames, were not part of our intended data format, adding another layer of complexity to our data selection process since most datasets did not provide video data.

Given these constraints, we ended up using both the TS-WorldCup Dataset [17] in an initial phase and later the SoccerNet dataset [34] [19]. The TS-WorldCup Dataset offered several beneficial features, such as homography ground truth for each frame, division into left and right field scenarios and short-length videos. The short length of the videos seemed to fulfil the processing, the costs of dealing with multiple video frames, and the computational costs inherent in optical flow processing. However, these short videos also posed a limitation. Since they were divided into left and right scenarios, we could not test and experiment with scenarios involving more prolonged camera movements, such as moving from one of the boxes to the middle of the field or vice versa.

This limitation led us to the SoccerNet dataset, which provided videos more suited to our needs. More specifically, the dataset that is provided for the Game State Recognition task which aims at recognising the state of a sport game, by identifying and localizing all sports individuals. Even though the ground truth did not have the homography, we had detailed information about players and the referee positions in the frame and in the pitch model, which was extremely helpful in further mapping evaluation. Another consideration was selecting the videos with as few zoom changes as possible. Even though broadcast cameras are usually PTZ cameras, we focused more on understanding the pan movements than other camera movements. Tilt movements are less prevalent in football videos, and in instances of a zooming action, the average motion may indeed approach zero, particularly if the zoom is centered. This phenomenon occurs because the motion vectors on one side of the frame will be in one direction, while the motion vectors on the opposite

side will be in the reverse direction. When these vectors are averaged, they could potentially negate each other, resulting in an average motion vector that is nearly zero.

To fulfil the available ground truth, we incorporated motion information from the optical flow and the MPEG-encoded motion vectors. By integrating the motion information from both methods into our ground truth data, we facilitated easy access to our calculations and methodology. Lastly, and since we were not provided with the ground truth homography for each frame, we needed the homography for the first frame to be able to predict how the homography changes according to the motion of the camera but also at least a reasonable number of ground truth homographies for evaluation purposes, so we manually computed it using the manual homography implementation.

For testing and validation purposes, it was important that in the process of that selection, we had a diverse range of scenarios. For instance, we included videos where the camera moved from the centre of the field towards one of the goal areas and vice versa. This selection strategy ensured that we tested both leftward and rightward movements. Additionally, we selected videos that featured mixed and opposite panning movements within the same sequence. This selection strategy ensured that our tests covered cases with repeated information and cases where the panorama grows to the left and right relative to the first frame, which is crucial for creating accurate panoramas. The following table presents a concise description of each video used for testing. It includes the duration of the video but also the relevant camera motion or the part of the pitch viewed along the video.

Table 5.1: Video's Description

Video	**Path Description (Camera Movement Direction)**	**Duration**
1	From left goal area to right goal area (right)	0:19
2	From centre circle to right goal area (right)	0:08
3	From right goal area to centre circle (left)	0:12
4	From left goal area to right goal area and back to left goal area (mixed)	0:21
5	From centre circle to left goal area (left)	0:08

Our experiments were conducted using an NVIDIA GeForce GTX 1050 Mobile graphics card, which also supported state-of-the-art models with some adaptations to parameters, such as reducing the original batch size and number of workers to ensure successful execution. Despite the relatively robust specifications, computing optical flow proved too resource-intensive for this setup. To overcome this limitation, we leveraged Google Colab's available TPUv2, which provided the necessary computational power to run the optical flow algorithms efficiently. This adjustment enabled us to perform the required computations without compromising performance or accuracy.

Chapter 6

Results

This chapter presents the results of our homography calculations for football matches captured using moving cameras.

Following this, subsequent sections detail the evaluation process of our experiments. This includes a comprehensive comparison of our methods against state-of-the-art benchmarks, the various techniques employed for calculating camera motion, and the different approaches used for homography calculation.

To better understand our findings, we will present quantitative metrics from the experimental setup and visual outcomes of our results. Simultaneously, we will scrutinise, analyse, and discuss these findings in-depth, focusing on the results and the comparisons to highlight the strengths and weaknesses.

6.1 State-of-the-Art Methods Evaluation

This section provides a comprehensive analysis of the results of the experiments applied to the state-of-the-art methods. We will focus on how these methods handle frames with insufficient keypoints.

The TVCalib approach showed strong performance across both key metrics evaluated. It consistently produced low error rates, demonstrating its reliability and robustness in various scenarios. This can be observed in the IoU results in Table 6.1, which display an overall comparison of the IoU for the two benchmark methods tested.

IoU Mean		
	TVCalib	Sportsfield Release
Video 1	0.97	0.90
Video 2	0.92	0.82
Video 3	0.91	0.83
Video 4	0.91	0.69
Video 5	0.97	0.82

Table 6.1: IoU Results of the State-of-the-Art methods

However, during our analysis, we encountered some outliers that slightly deviated from the expected performance range. To illustrate this, we have included a histogram in Figure 6.1, showing the error distribution for one of the videos we analysed. This histogram demonstrates the overall effectiveness of the TVCalib method, but it also highlights the presence of a few outliers.

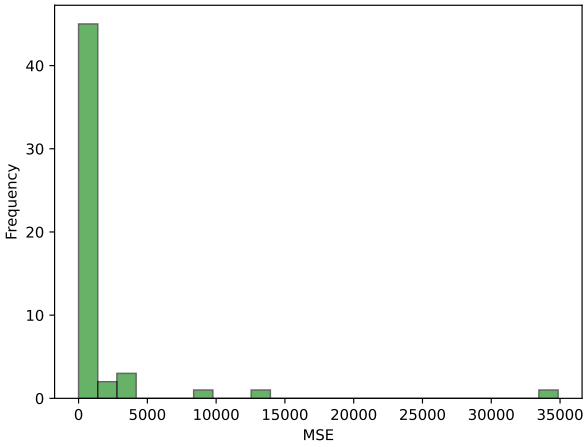


Figure 6.1: MSE Histogram - TVCalib Video 4

Upon investigating the outliers, we found that most of them originated from frames with few keypoints. Considering that this video encompasses the entire field, it naturally becomes more susceptible to finding zones with few keypoints. Some examples of these frames are represented in the following images (Figure 6.2).



(a) Frame with few keypoints leading to outlier

(b) Another example of an outlier with reduced keypoints

Figure 6.2: Examples of TVCalib Outliers

The blurriness in the images could be caused by fast camera movements, suggesting that the camera may not handle rapid motions compared to minor movements in other parts of the video. This imprecision could be attributed to the frame-by-frame prediction method used instead of a motion-based or a temporal consistent processing. This shows some limitations in dealing with a moving camera by not handling homography well in fast motion frames.

The SportsField approach is another state-of-the-art method that was evaluated in our study. Based on IoU metrics, it did not achieve comparable performance to TVCalib, as indicated in

Table 6.1. Specifically, when analysing the results of the previous video, Video 4, with a significant number of frames containing few keypoints, the SportsField approach exhibited even poorer results for both IoU and MSE, as shown in Figure 6.3.

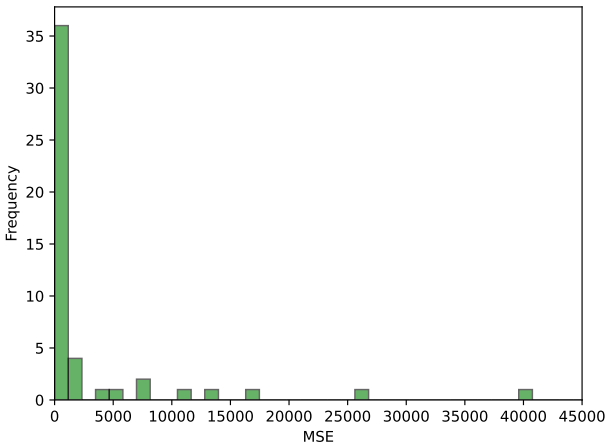


Figure 6.3: MSE Histogram - Sportsfield Video 4

This disparity in results between the two methods could potentially be attributed to the diversity and volume of the training data available. TVCalib emphasises the advantage of a more comprehensive training dataset that covers a variety of scenarios and camera types. In contrast, the SportsField approach is tailored to a more specific dataset, focusing primarily on main broadcast cameras. This difference in the scope and depth of training data could significantly influence the performance of these two methods.

Furthermore, examining another video example highlighted that many of the outliers for this method were also associated with frames characterised by limited keypoints. This finding is consistent with the histogram shown in Figure 6.4, which illustrates the distribution of errors across frames.

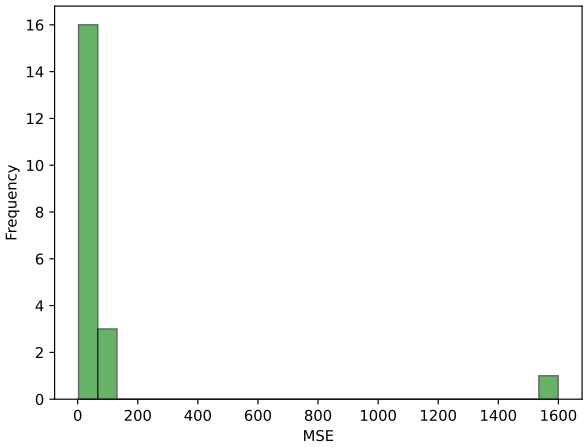


Figure 6.4: MSE Histogram - Sportsfield Video 2

By comparing the outlier frames with their respective images (Figure 6.5), it becomes evident once again that these frames contain few keypoints. Furthermore, the relatively low number of keypoints could also be attributed to occlusions caused by players. As represented in Figure 6.5a, there are numerous players positioned near the lines and corners of the football field, which may lead to partial occlusion of key areas.



(a) Example of Frame with Few Keypoints Leading to Outlier



(b) Example of Frame with Sparse Keypoints Contributing to Outlier

Figure 6.5: Examples of Sportsfield Outliers

This disparity highlights the methods' susceptibility to scenarios where the number of keypoints is reduced, impacting their accuracy in registering sports field templates and estimating the corresponding homography under such conditions.

6.2 Super Image Approach

In this upcoming analysis, we will start by exploring the super image approach by examining the IoU metric, which offers valuable insights into the homography calculation when employing this method. The comprehensive results of this metric can be found in the Table 6.2, below.

IoU Mean			
	Super Image		
	Manual	Optical Flow	MPEG
Video 1	0.86	0.87	0.81
Video 2	-	0.77	0.78
Video 3	-	0.66	0.65
Video 4	-	0.58	0.57
Video 5	-	0.73	0.71

Table 6.2: IoU Results of Super Image Approach

We began by using a manual approach as a baseline for our analysis. For the first video, the manual method produced results similar to those from MPEG and Optical Flow. Due to the exhaustive nature of manually selecting points for an entire video sequence, we evaluated only the Optical Flow and MPEG methods for the subsequent videos. Regarding these results, the IoU values are generally lower than those of state-of-the-art techniques. This can be explained with the help of the graph shown in Figure 6.6.

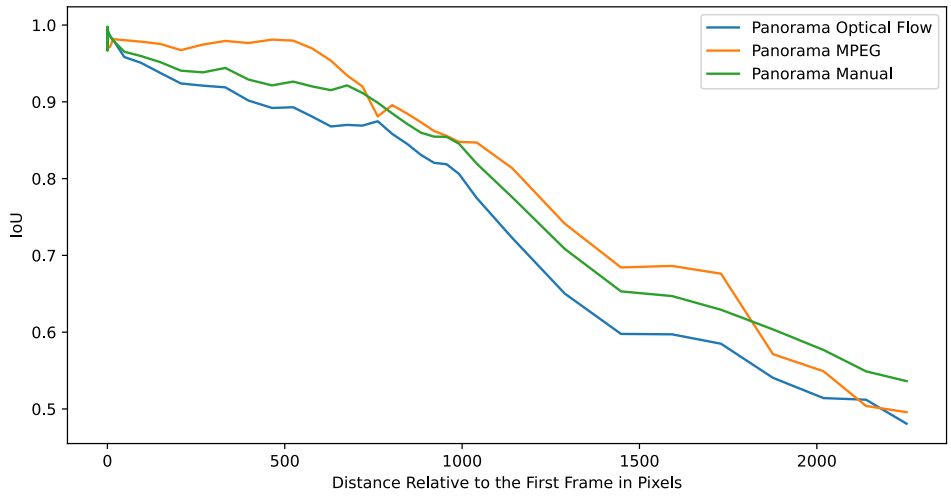
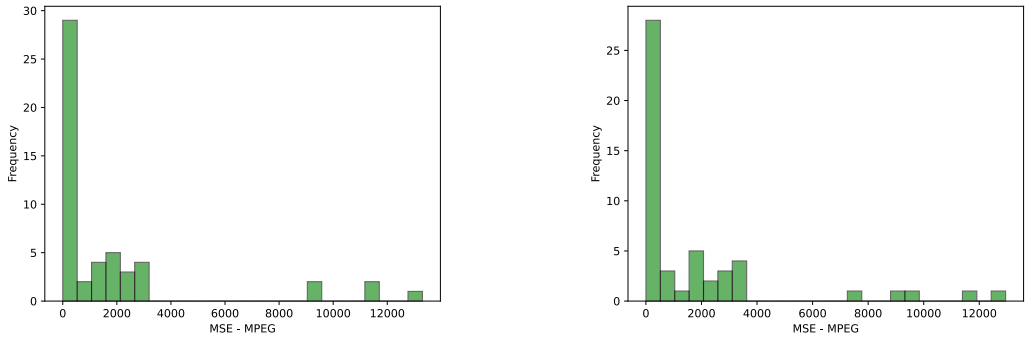


Figure 6.6: Evolution of IoU according to the distance relative to the first frame - Video 1

An interesting trend observed in all the approaches was the decrease in IoU over the course of the videos. Figure 6.6 shows the results from Video 1, using the three different motion calculation methods. The IoU varies based on the distance from the first frame. The farther the distance from the initial frame, the higher the error introduced, resulting in a high initial score for the first frames. Although, as the distance increases, the IoU decreases, leading to a lower overall mean score.

However, we observe that the introduced error is inferior when comparing the results obtained using these methods with frames characterised by insufficient keypoints from the state-of-the-art results. It seems that the introduced error is less significant when comparing the results obtained using these methods with frames that have a lack of keypoints, as observed in the state-of-the-art

results. For example, we can examine the histograms for the MSE evaluation for Video 4 for both the MPEG and Optical Flow implementations in Figure 6.7a and Figure 6.7b , respectively.



(a) MSE Histogram - MPEG

(b) MSE Histogram - Optical Flow

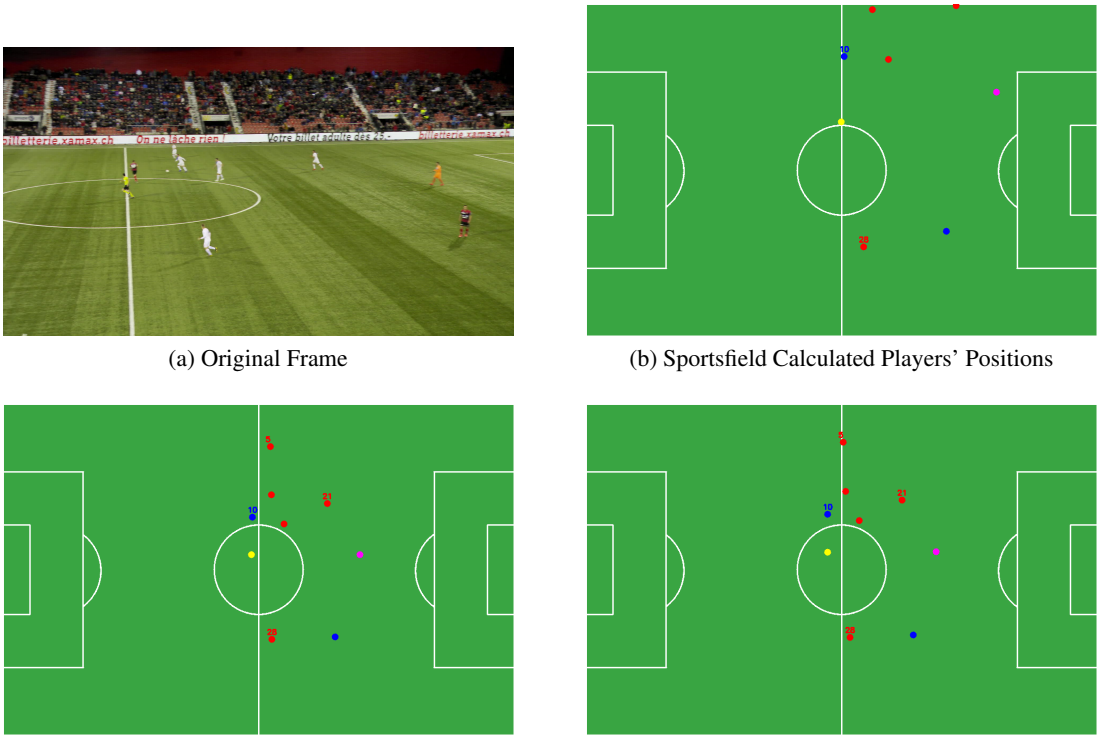
Figure 6.7: MSE Results of Video 4 Using the Super Image Approach

When we look at these results and compare them with the results from Figure 6.1, we notice that the errors in our methods are more evenly distributed than the TVCalib results. However, the outliers in our methods are less extreme in comparison. Therefore, despite our approach having increased error in camera displacement, the error is less significant when there are few keypoints in the frame.

When comparing outliers in the latest results, visually analysing players' positions can give us valuable insights. For instance, let's examine the predicted positions of players for an outlier using the Sportsfield approach in Video 5. Figure 6.8a presents the original frame, Figure 6.8b the Sportsfield calculated players' position based on the predicted homography, Figure 6.8c and Figure 6.8d the calculated positions based on the homography obtained using the Super Image approach for the MPEG and Optical Flow, respectively. For this visual result, we use a pitch template to display the predicted positions of players from both teams, along with additional key figures such as referees. The colour coding is the following:

- Team defending the left goal: Displayed as blue dots, with their goalkeeper represented by a cyan dot.
- Team defending the right goal: Depicted as red dots, with their goalkeeper denoted by a magenta dot.
- Referee: Shown as a yellow dot.

Certain players are identified with numbers based on the ground truth (GT) data available in the dataset. The presence of a player's number indicates that this specific information is contained in the dataset GT. This labelling simplifies the analysis of our results by clearly identifying individual players. In the following example, the blue team represents players in black jerseys, while the red team represents players in white jerseys.



(a) Original Frame

(b) Sportsfield Calculated Players' Positions

(c) Super Image Approach (MPEG)

(d) Super Image Approach (Optical Flow)

Figure 6.8: Comparison of Predicted Players' Positions for an Outlier in Video 5

Upon analysing these images, it is clear that the Sportsfield approach has the worst prediction when compared to the original frame. The positions of the players are less accurate and more dispersed, highlighting the limitations of the Sportsfield method in this scenario.

We have already mentioned that the manual approach showed results similar to those of the two automated methods. We will now compare the MPEG and Optical Flow approaches directly. Analysing Table 6.2 and Figure 6.7, we can see that the results were also similar. Let us compare two panoramas created using different methods for Video 3.



Figure 6.9: Panorama Video 3 - Optical Flow



Figure 6.10: Panorama Video 3 - MPEG

In Figure 6.9 we can see the panorama created using Optical Flow motion data, while Figure 6.10 shows the results for the same video, but using MPEG encoded motion. Additional results from other videos can be found in Appendix A. The accuracy of these panoramas appears to be contingent on the camera movement, with some methods yielding more precise results than others. The precision of the generated panoramas seems to be intricately linked to the nature of the camera movement. The analysis indicates that factors such as the level of zoom and the degree of vertical displacement could significantly impact the accuracy of the panorama. The analysis of the results indicates that both the Optical Flow and MPEG-derived motion vectors yield similar levels of accuracy in the generation of video panoramas and in the calculated metrics. Interestingly, the use of MPEG can offer results that are comparable to those of the Optical Flow method while potentially reducing time and computational costs due to the motion supplied by the MPEG streams. Additionally, motion analysis can be conducted without full-frame decoding, providing another cost advantage.

To conclude the analysis of the Super Image approach it is important to refer that even if this particular approach may not achieve the highest scores numerically, it can still provide valuable qualitative insights and help in situations where it is difficult to find correspondences to calculate the homography. These insights can be crucial for game analysis and tactical evaluation. In some cases, they offer a unique perspective compared to state-of-the-art techniques. For example, by mapping a specific frame (Figure 6.11) on the generated panorama, we can locate the frame spatially, as seen in Figure 6.12



Figure 6.11: Original Frame



Figure 6.12: Example of a Frame mapped in a Super Image generated using MPEG motion vectors

6.3 Equations Approach

In this section, we will analyse our second approach, which directly utilises motion in the homography equations. This method involves deriving motion data from manual inputs, optical flow, and MPEG data, and then applying these motions directly to the homography transformation equations.

Upon conducting an analysis of the IoU metrics from Table 6.3, it is observed that the results were marginally lower in comparison to those obtained using the Super Image methodology, though they remained closely aligned in overall performance.

IoU Mean			
	Equations		
	Manual	Optical Flow	MPEG
Video 1	0.87	0.87	0.84
Video 2	-	0.74	0.74
Video 3	-	0.65	0.56
Video 4	-	0.53	0.53
Video 5	-	0.74	0.66

Table 6.3: IoU Results of Equations Approach

A similar trend was observed in both the Equations and Panorama approaches, where the IoU tended to decrease with continuous camera movement. This indicates that both methodologies were sensitive to camera motion, as the IoU values fluctuated accordingly, demonstrating resilience in maintaining accuracy across multiple frames. Yet in this approach we found that it increased when the camera returned to a previous position or changed its movement direction to get closer to the original frame. These findings are represented in Figure 6.13.

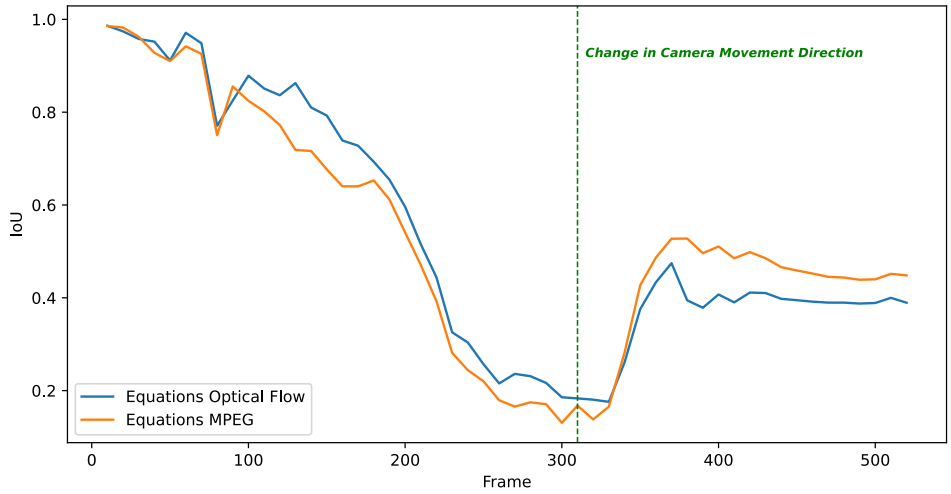
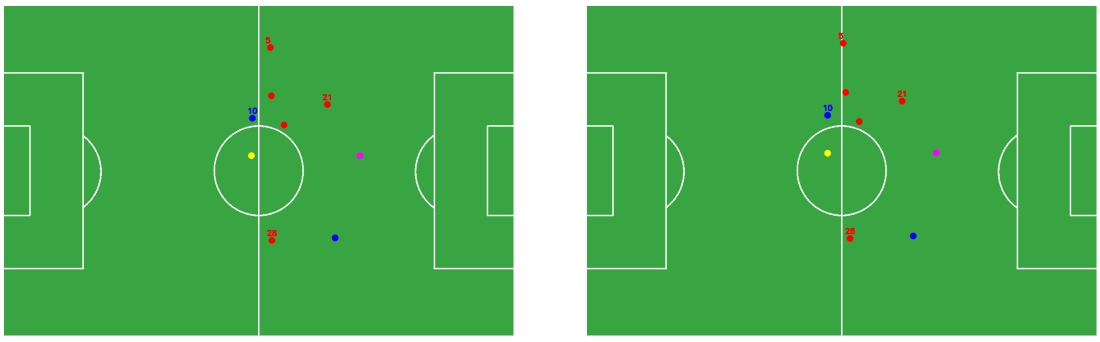


Figure 6.13: Evolution of IoU over the video frames

Figure 6.13 represents a situation in which the camera direction remains consistent until approximately frame 300. During this period, the IoU consistently decreases. After frame 300, the camera changes direction and moves closer to the initial starting point, which results in a slight increase in the IoU. This behaviour demonstrates how the approach is sensitive to continuous camera movement and its capacity to regain accuracy when the camera's movement aligns more closely with the original position.

Comparing the MSE results to the other approach, we can see that, in general, the results were very similar. For instance, let us look at the calculated player's positions using the homography obtained from this Equations approach and compare it to the results from Figure 6.8. As shown in Figure 6.14, we got similar player positions from the Super Image approach.

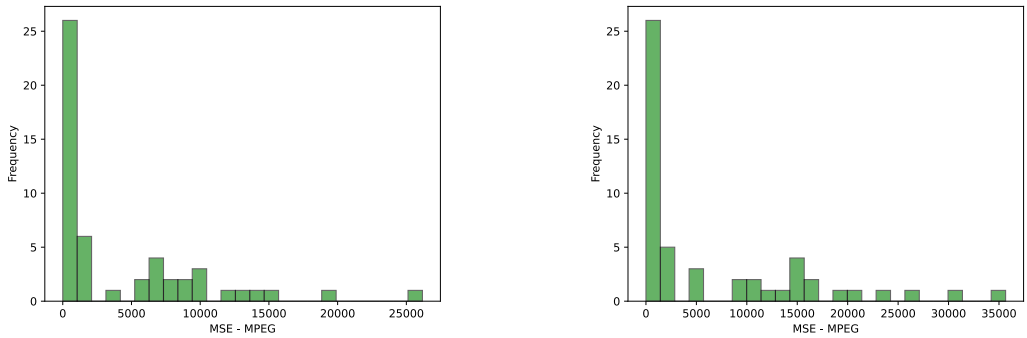


(a) Equations Approach (MPEG)

(b) Equations Approach (Optical Flow)

Figure 6.14: Comparison of Predicted Players' Positions for an Outlier in Video 5

To provide a more quantitative analysis, we can compare the Equations approach with the Super Image results, shown in Figure 6.7. By examining the histogram, we can better understand how the Equations approach performs relative to the Super Image method in terms of MSE.



(a) MSE Histogram - MPEG

(b) MSE Histogram - Optical Flow

Figure 6.15: MSE Results of Video 4 Using the Super Image Approach

Upon observing the histogram, it is evident that the Equations approach produced less favorable results, with errors that were more widely distributed and with more extreme outliers compared to the Super Image method. This suggests that the Equations approach was overall less effective. Furthermore, the incorporation of the panorama in the Super Image approach provided a more comprehensive visual dataset compared to the Equations approach. Variations in the selection and use of keypoints across different videos may explain the inconsistencies observed in the Equations approach. This variability could account for why some results were comparable to those of the Super Image approach while others were notably worse.

Chapter 7

Final Discussion

This concluding chapter summarises our investigation into calculating homography in football matches under moving camera scenarios. We will outline the main findings from our suggested methods and existing techniques, highlighting the advantages and limitations identified during our experiments. Furthermore, we will propose enhancements to the approach based on the identified problems.

7.1 Conclusions

In our research, we focused on improving the calculation of homography in dynamic football match scenarios. We aimed to address the limitations of current methods in handling frames with insufficient keypoints and ensuring temporal consistency. We evaluated different approaches for comparison. To address these issues, we proposed using motion information obtained from various inputs. We compared a manual approach serving as a baseline, optical flow, and encoded motion vectors from MPEG video coding.

In this study, we utilised two methods to analyse this. Firstly, TVCalib consistently yielded low error rates and high IoU scores, demonstrating its robustness across various scenarios. However, it faced challenges with rapid camera movements and frames with few keypoints. For example, blurred images resulting from fast movements exhibited the poorest results with this approach, as did frames near the midfield, which had fewer corners and edges compared to the goal area. This was even more evident in frames where the centre circle was only partially visible. The second technique, SportsField, also struggled in scenarios with insufficient keypoints in those same zones of the pitch. While it showed some robustness, its performance was significantly impacted in videos with a few keypoints correspondences. We also notice that even if there were parts of the pitch visible, they might be partially occluded by players producing undesired results. Finally, this last approach was generally worse than the first, probably due to the nature of the data these models were trained with. The first model, TVCalib, was trained on a more extensive dataset that included various cameras and points of view. In contrast, the second model was trained predominantly on

the main broadcast camera frames. This discrepancy in training data could explain the differences in performance on unseen data since we used a dataset that was not used in these models' training.

In our proposed approach, we utilised two distinct techniques. The equations approach involved the direct integration of motion data into the homography equations, enabling us to update the homography dynamically based on the detected motion. The super image approach leveraged motion data to generate a panorama, offering a comprehensive visual representation of the field. This motion was then utilised to adjust the homography of the first frame, thus contributing to more accurate calculations of the subsequent frames' homography.

After evaluating the super image approach, we found that its performance metrics generally show lower accuracy compared to state-of-the-art methodologies. However, it was observed that in situations with fewer keypoints, the super image approach demonstrated less error compared to state-of-the-art techniques. Our study also revealed that the proposed methods maintain their accuracy for an initial set of frames. Still, as the camera movement persists, moving away from the starting point, this accuracy tends to diminish. This suggests that the approach could help manage sporadic occlusions or temporary lack of keypoints during video sequences. Nevertheless, this approach may not be suitable as the sole method for extended sequences due to decreased performance over the video as it progresses away from the starting point.

It is essential to note that qualitative observations are equally significant for the panoramic visual outputs, which exhibited favourable results. Even if specific methods may not yield the highest quantitative scores, they can still provide valuable qualitative insights. These insights can be crucial for game analysis and tactical evaluation, and in some cases, they offer a unique perspective compared to cutting-edge techniques. For instance, mapping a frame onto the created panorama provides a comprehensive view of player positions and movements across the entire pitch, allowing for better strategic reviews and identifying patterns or anomalies in team formations and play styles. Moreover, the precision of the generated panoramas is closely tied to the nature of the camera movement. The analysis suggests that factors such as the level of zoom and the degree of vertical displacement could have a significant impact on the accuracy of the panorama. For instance, the third video, characterised by minimal vertical displacement, seems to produce the most accurate panorama.

Observations reveal parallel trends between the equations and panorama approaches. The equations approach also demonstrates resilience in maintaining accuracy across multiple frames, but it also revealed less consistency across the videos tested compared to the super image results. This inconsistency was likely due to the variability in the selected points used to calculate the homography, which produced different results across various videos. However, an interesting observation in this approach was that the error tends to decrease when a camera shift makes the frame return closer to the initial frame zone of the pitch.

Comparing the different motion inputs, the manual approach used as a benchmark provided a high-confidence reference for evaluating other approaches. However, it was impractical for extensive video sequences due to its exhaustive nature. In both methods, the analysis of the findings indicates that optical flow and MPEG-derived motion vectors produce similar levels of accuracy in

obtained homographies and in creating video panoramas. Notably, employing MPEG can deliver results comparable to the optical flow approach. Since MPEG data is readily available in compressed video streams, it eliminates computational expenses linked to the optical flow estimations. This makes MPEG suitable for resource-constrained environments due to its greater efficiency compared to optical flow techniques while keeping reliability, as the derived motion vectors are part of the video compression process and are inherently tied to the video content.

7.2 Future Work

Current methods face challenges in accurately aligning frames during continuous camera movements, which can lead to decreased alignment over time. It is essential to improve the integration of motion data into homography calculations. Tackling these challenges will be crucial for advancing the strength and dependability of motion-based homography techniques in dynamic video environments. Moreover, enhancing alignment accuracy along the y-axis and developing strategies to manage zoom variations could reduce inaccuracies in homography calculations, especially in the visual results of panoramas.

Applying AI and deep learning techniques offers significant potential for advancing homography calculation methods. Future studies could explore integrating motion information as inputs into state-of-the-art models, thereby improving their capability to predict precise homographies in challenging scenarios. Training these models on diverse datasets encompassing various camera viewpoints and motion patterns could enhance their generalisation and effectiveness across different sports scenarios. Another strategy could be hybrid approaches combining state-of-the-art methods with motion techniques. For instance, integrating motion-based corrections for frames where existing methods show reduced performance could potentially enhance overall accuracy. This hybrid approach could involve selectively applying corrections to specific frames or video segments to validate accuracy improvements.

This future research focus could also open the way for advancements in real-time applications for sports analysis. Developing efficient methods to ensure high accuracy in homography calculations with a moving camera would allow for processing live video feeds in real time. This capability would be crucial for sports production teams, enabling instant analysis of player positions and tactical movements, as well as interactive visualisations during live broadcasts.

References

[1] Habtamu Zegeye Alemu, Wei Wu, and Junhong Zhao. Feedforward neural networks with a hidden layer regularization method. *Symmetry*, 10(10), 2018.

[2] M. Alemán-Flores, L. Alvarez, L. Gomez, P. Henriquez, and L. Mazorra. Camera calibration in sport event scenarios. *Pattern Recognition*, 47(1):89–95, 2014.

[3] James A Anderson. *An introduction to neural networks*. MIT press, 1995.

[4] Supavadee Aramvith and Ming-Ting Sun. Chapter 9 - mpeg-1 and mpeg-2 video standards. In Al Bovik, editor, *The Essential Guide to Video Processing*, pages 267–293. Academic Press, Boston, 2009.

[5] The Football Association. Law 1 - the field of play, 2023. FA. Available at https://www.thefa.com/football-rules-governance/lawsandrules/laws/football-11-11/law-1---the-field-of-play, last accessed in January 2024.

[6] The Football Association. Law 1 the field of play, 2023. IFAB. Available at https://www.theifab.com/laws/latest/the-field-of-play/#field-surface, last accessed in January 2024.

[7] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Comput. Surv.*, 27(3):433–466, sep 1995.

[8] Daniel Berjón, Carlos Cuevas, and Narciso García. Soccer line mark segmentation and classification with stochastic watershed transform. *Signal Processing: Image Communication*, 118:117014, 2023.

[9] Chris M Bishop. Neural networks and their applications. *Review of scientific instruments*, 65(6):1803–1832, 1994.

[10] Erin Blakemore. Soccer is the world's most popular sport. but who invented it?, 2023. National Geographic. Available at https://www.nationalgeographic.com/premium/article/soccer-world-cup-origins-mesoamerica-football-games-archaeology, last accessed in November 2023.

[11] Andre Bleau and L Joshua Leon. Watershed-based segmentation and region merging. *Computer Vision and Image Understanding*, 77(3):317–370, 2000.

[12] L. Bommes, X. Lin, and J. Zhou. Mvmed: Fast multi-object tracking in the compressed domain. In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1419–1424, 2020.

[13] Afshin Bozorgpour, Mehran Fotouhi, and Shohreh Kasaei. Robust homography optimization in soccer scenes. In *2015 23rd Iranian Conference on Electrical Engineering*, pages 787–792. IEEE, 2015.

[14] Christopher Brumann and Markus Kukuk. Evolution based single camera resectioning based on distance maps of a known geometry for squash sports. *IEEE Access*, 10:58136–58150, 2022.

[15] Jianhui Chen and James J Little. Sports camera calibration via synthetic data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[16] Yi Chen, Jun Bin, and Chao Kang. Application of machine vision and convolutional neural networks in discriminating tobacco leaf maturity on mobile devices. *Smart Agricultural Technology*, 5:100322, 2023.

[17] Yen-Jui Chu, Jheng-Wei Su, Kai-Wen Hsiao, Chi-Yu Lien, Shu-Ho Fan, Min-Chun Hu, Ruen-Rone Lee, Chih-Yuan Yao, and Hung-Kuo Chu. Sports field registration via keypoints-aware label condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3523–3530, June 2022.

[18] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific data*, 9(1):355, 2022.

[19] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3502, 2022.

[20] P Dalka. *Methods of algorithmic analysis of the video image for applications in traffic monitoring*. PhD thesis, Ph. D. thesis, Gdansk University of Technology, 2015.

[21] Elan Dubrofsky. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5, 2009.

[22] Karol Działowski and Paweł Forczmański. Football players movement analysis in panning videos. In Maciej Paszynski, Dieter Kranzlmüller, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M.A. Sloot, editors, *Computational Science – ICCS 2021*, pages 193–206, Cham, 2021. Springer International Publishing.

[23] Hazem Elbakry Ebtsam Adel, Mohammed Elmogy. Image stitching based on feature extraction techniques: A survey. *International Journal of Computer Applications*, 99(6):1–8, August 2014.

[24] Ramon L. Felip, LluÍs Barcelo, Xavier Binefa, and John R. Kender. Robust dominant motion estimation using mpeg information in sport sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(1):12–22, 2008.

[25] FIFA. Goal-line technology. FIFA. Available at https://www.fifa.com/technical/football-technology/standards/goal-line-technology, last accessed in November 2023.

[26] FIFA. Fifa semi automated offside technology animation, 2022. FIFA. Available at https://www.youtube.com/watch?v=FEPKAeGzwPU, last accessed in November 2023.

[27] FIFA. Goal-line technology, 2022. FIFA. Available at https://www.fifa.com/technical/football-technology/football-technologies-and-innovations-at-the-fifa-world-cup-2022/goal-line-technology, last accessed in November 2023.

[28] FIFA. Semi-automated offside technology, 2022. FIFA. Available at https://www.fifa.com/technical/football-technology/football-technologies-and-innovations-at-the-fifa-world-cup-2022/semi-automated-offside-technology, last accessed in November 2023.

[29] FIFA. Semi-automated offside technology to be used at fifa world cup 2022™, 2022. FIFA. Available at https://www.fifa.com/technical/media-releases/semi-automated-offside-technology-to-be-used-at-fifa-world-cup-2022-tm, last accessed in November 2023.

[30] FIFA. 5.3 pitch dimensions and surrounding areas, 2023. FIFA. Available at https://publications.fifa.com/en/football-stadiums-guidelines/technical-guideline/stadium-guidelines/pitch-dimensions-and-surrounding-areas/, last accessed in January 2024.

[31] Mehran Fotouhi, Afshin Bozorgpour, and Shohreh Kasaei. Automatic soccer field line recognition by minimum information. In *Artificial Intelligence and Signal Processing (AISP), 2015 International Symposiumon*, pages 136–142. IEEE, 2015.

[32] Mehran Fotouhi, Sadjad Fouladi, and Shohreh Kasaei. Projection matrix by orthogonal vanishing points. *Springer, Multimedia Tools and Applications*, 76(15):16189–16223, August 2017.

[33] Mengyin Fu, Hao Liang, Chunhui Zhu, Zhipeng Dong, Rundong Sun, Yufeng Yue, and Yi Yang. Image stitching techniques applied to plane or 3-d models: A review. *IEEE Sensors Journal*, 23(8):8060–8079, 2023.

[34] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkin, Ning Wang, Qiong Jia, Quoc Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Masui, Shouhong Ding, Sin wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. SoccerNet 2022 challenges

results. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*. ACM, oct 2022.

[35] Google. What is deep learning?, 2023. Google Cloud. Available at https://cloud.google.com/discover/what-is-deep-learning, last accessed in December 2023.

[36] Xiaoming Guo, Guang Jiang, Zhaopeng Cui, and Pei Tao. Homography-based block motion estimation for video coding of ptz cameras. *Journal of Visual Communication and Image Representation*, 39:164–171, 2016.

[37] Marc Gutiérrez-Pérez and Antonio Agudo. No bells, just whistles: Sports field registration by leveraging geometric properties, 2024.

[38] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[39] IBM. What is deep learning?, 2023. IBM. Available at https://www.ibm.com/topics/deep-learning, last accessed in December 2023.

[40] IBM. What is machine learning?, 2023. IBM. Available at https://www.ibm.com/topics/machine-learning, last accessed in December 2023.

[41] Wei Jiang, Juan Camilo Gamboa Higuera, Baptiste Angles, Weiwei Sun, Mehrsan Javan, and Kwang Moo Yi. Optimizing through learned errors for accurate sports field registration, 2020.

[42] Chetan G Kagalagomb and Sunanda Dixit. Tracking of soccer players using optical flow. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 2*, pages 117–129. Springer, 2021.

[43] Stephen Karungaru, Hiroki Tanioka, and Kenji Matsuura. Soccer players real location determination using perspective transformation. In *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pages 1–4, 2022.

[44] John D Kelleher. *Deep learning*. MIT press, 2019.

[45] Arati Kushwaha, Ashish Khare, Om Prakash, and Manish Khare. Dense optical flow based background subtraction technique for object segmentation in moving camera environment. *IET Image Processing*, 14(14):3393–3404, 2020.

[46] ANGIE LEE. Kicking games up a notch: Startup sports vision ai to broadcast athletics across the globe, 2023. NVIDIA Blogs. Available at https://blogs.nvidia.com/blog/pixellot-vision-ai-sports-broadcasting/, last accessed in November 2023.

[47] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019, 2022.

[48] Deutsche Fussball Liga. Track160. Deutsche Fussball Liga. Available at https://www.dfl.de/en/glossary/track160/, last accessed in November 2023.

[49] La Liga. How laliga's mediacoach allows teams to compete using data, 2021. La Liga Newsletter. Available at https://newsletter.laliga.es/global-futbol/laliga-mediacoach-clubs-compete-using-data, last accessed in November 2023.

[50] La Liga. How laliga's mediacoach platform brings value to clubs, broadcasters and fans, 2021. La Liga Newsletter. Available at https://newsletter.laliga.es/global-futbol/laliga-mediacoach-platform-brings-value-to-clubs-broadcasters-and-fans, last accessed in November 2023.

[51] La Liga. Laliga and microsoft create 'beyond stats', a statistics system that enhances the fan experience thanks to big data, 2021. La Liga Newsletter. Available at https://newsletter.laliga.es/global-futbol/laliga-and-microsoft-create-beyond-stats-a-statistics-system-that-enhances-the-fan-experience-thanks-to-big-data, last accessed in November 2023.

[52] La Liga. Laliga paves the way for the future of bi and analytics in football thanks to mediacoach and the beyond stats project, 2023. La Liga Press Release. Available at https://www.laliga.com/en-GB/news/laliga-paves-the-way-for-the-future-of-bi-and-analytics-in-football-thanks-to-mediacoach-and-the-beyond-stats-project, last accessed in November 2023.

[53] Chung-Ching Lin, Sharathchandra U Pankanti, Karthikeyan Natesan Ramamurthy, and Aleksandr Y Aravkin. Adaptive as-natural-as-possible image stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1163, 2015.

[54] Antje Linnemann, Sebastian Gerke, Sven Kriener, and Patrick Ndjiki-Nya. Temporally consistent soccer field registration. In *2013 IEEE International Conference on Image Processing*, pages 1316–1320, 2013.

[55] Yinhui Luo, Xingyi Wang, Yanhao Liao, Qiang Fu, Chang Shu, Yuezhou Wu, and Yuanqing He. A review of homography estimation: Advances and challenges. *Electronics*, 12(24), 2023.

[56] Adrien Maglo, Astrid Orcesi, Julien Denize, and Quoc Cuong Pham. Individual locating of soccer players from a single moving view. *Sensors*, 23(18):7938, 2023.

[57] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Kalicalib: A framework for basketball court registration. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, pages 111–116, 2022.

[58] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.

[59] Mehrtash Manafifard. A review on camera calibration in soccer videos. *Multimedia Tools and Applications*, pages 1–32, 2023.

[60] RICK MERRITT. Ai on the ball: Startup shoots computer vision to the soccer pitch, 2022. Blogs NVIDIA. Available at https://blogs.nvidia.com/blog/ai-soccer-track160/, last accessed in November 2023.

[61] Microsoft. Shaping the future of the game. Microsoft Unlocked. Available at https://unlocked.microsoft.com/laliga-beyond-stats/, last accessed in November 2023.

[62] Berndt Müller, Joachim Reinhardt, and Michael T Strickland. *Neural networks: an introduction*. Springer Science & Business Media, 1995.

[63] Banoth Thulasya Naik, Mohammad Farukh Hashmi, and Neeraj Dhanraj Bokde. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Applied Sciences*, 12(9), 2022.

[64] NBC. 'premier league data zone' enhanced viewing experience debuts exclusively on peacock for this sunday's arsenal-liverpool match, 2023. NBC Press Release. Available at https://nbcsportsgrouppressbox.com/2023/04/06/premier-league-data-zone-enhanced-viewing-experience-debuts-exclusively-on-peacock-for-this-sundays-arsenal-liverpool-match/, last accessed in November 2023.

[65] Xiaohan Nie, Shixing Chen, and Raffay Hamid. A robust and efficient framework for sports-field registration. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1935–1943, 2021.

[66] OpenCV. Optical flow, 2023. OpenCV. Available at https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html, last accessed in May 2024.

[67] Yash Pandya, Kaustav Nandy, and Shivam Agarwal. Homography based player identification in live sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5208–5217, 2023.

[68] Jung Me Park and Yi Lu Murphey. Edge detection in grayscale, color, and range images. *Wiley Encyclopedia of Computer Science and Engineering*, pages 1–16, 2007.

[69] Sanskruti Patel. An overview and application of deep convolutional neural networks for medical image segmentation. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pages 722–728, 2023.

[70] Stats Perform. 'revolutionise sports through ai'. Stats Perform. Available at https://www.statsperform.com/artificial-intelligence/, last accessed in December 2023.

[71] Stats Perform. 'football', 2023. Stats Perform Football. Available at https://www.statsperform.com/opta-football/, last accessed in December 2023.

[72] Stats Perform. 'opta data', 2023. OPTA DATA. Available at https://www.statsperform.com/opta/, last accessed in December 2023.

[73] Stats Perform. 'superior insight on both sides of the ball during live matches', 2023. OPTA VISION. Available at https://www.statsperform.com/opta-vision/, last accessed in December 2023.

[74] Vincenzo Pesce, Pablo Hermosin, Aureliano Rivolta, Shyam Bhaskaran, Stefano Silvestrini, and Andrea Colagrossi. Chapter nine - navigation. In Vincenzo Pesce, Andrea Colagrossi, and Stefano Silvestrini, editors, *Modern Spacecraft Guidance, Navigation, and Control*, pages 441–542. Elsevier, 2023.

[75] Pixellot. Automated soccer tracking camera. Pixellot. Available at https://www.pixellot.tv/sports/soccer/, last accessed in November 2023.

[76] Pixellot. Company. Pixellot. Available at https://www.pixellot.tv/about-us/about/, last accessed in November 2023.

[77] Pixellot. Introducing show s3: Reinventing live sports streaming. Pixellot. Available at https://www.pixellot.tv/blog/introducing-show-s3-reinventing-live-sports-streaming/, last accessed in November 2023.

[78] Diogo Pontes, Claudino Costa, Ricardo Gomes Faria, Pedro Fidalgo, and José Henrique Brito. Piracy detection in online soccer streaming with video content inspection: an application to the portuguese market. In *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–5, 2023.

[79] Genius Sports Press. Genius sports acquires second spectrum, the official data tracking and analytics provider of the epl, nba, and mls, 2021. Investors Genius Sports. Available at https://geniussports.com/newsroom/sportstech/genius-sports-acquires-second-spectrum-the-official-data-tracking-and-analytics-provider-of-the-epl-nba-and-mls/, last accessed in November 2023.

[80] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[81] Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, and Sujoy Ganguly. End-to-end camera calibration for broadcast videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13627–13636, 2020.

[82] Xi Shen, François Darmon, Alexei A Efros, and Mathieu Aubry. Ransac-flow: generic two-stage image alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 618–637. Springer, 2020.

[83] Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir Mohammad Mansourian, Xin Zhou, Shohreh Kasaei, Bernard Ghanem, Alexandre Alahi, Marc Van Droogenbroeck, and Christophe De Vleeschouwer. Soccernet game state reconstruction: End-to-end athlete tracking and identification on a minimap, 2024.

[84] Second Spectrum. Second spectrum chooses aws for cloud, machine learning, artificial intelligence, and media services. Second Spectrum. Available at https://www.secondspectrum.com/aws/, last accessed in November 2023.

[85] Second Spectrum. Win more games. Second Spectrum - Our Work. Available at https://www.secondspectrum.com/ourwork/teams-leagues, last accessed in November 2023.

[86] Spiideo. Soccer football video analysis software. Spiideo. Available at https://www.spiideo.com/soccer-football-video-analysis-software/, last accessed in November 2023.

[87] Spiideo. Spiideo play. Spiideo. Available at https://www.spiideo.com/spiideo-play/, last accessed in November 2023.

[88] Genius Sports. Genius sports tech. Genius Sports. Available at https://geniussports.com/sportstech/, last accessed in November 2023.

[89] BBC Sportshour. How ai is helping football, 2021. Sportshour Series. Available at https://www.bbc.co.uk/programmes/p09sfxm7, last accessed in November 2023.

[90] Richard Szeliski. *Introduction*, pages 1–26. Springer International Publishing, Cham, 2022.

[91] Kazumoto Tanaka. Karate video reconstruction to observe player motion and flow of competition. *International Journal on Advanced Science, Engineering and Information Technology*, 12(4):1344–1350, 2022.

[92] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 402–419, Cham, 2020. Springer International Publishing.

[93] Jonas Theiner and Ralph Ewerth. Tvcalib: Camera calibration for sports field registration in soccer, 2022.

[94] Graham Thomas. Real-time camera tracking using sports pitch markings. *Journal of Real-Time Image Processing*, 2(2):117–132, 2007.

[95] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18, 2017. Computer Vision in Sports.

[96] Track160. Track160 coach160. Track160. Available at https://www.track160.com/coach160, last accessed in November 2023.

[97] Track160. Track160 soccer technology. Track160. Available at https://www.track160.com/soccer-technology, last accessed in November 2023.

[98] Hiroki Tsurusaki, Keisuke Nonaka, Ryosuke Watanabe, Tomoaki Konno, and Sei Naito. [paper] sports camera calibration using flexible intersection selection and refinement. *ITE Transactions on Media Technology and Applications*, 9(1):95–104, 2021.

[99] Sruthi Viswanathan, Ravi Shankar, Jack Jackson, and Reuben Binns. Spectators of ai: Football fans vs. the semi-automated offside technology at the 2022 fifa world cup. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2023.

[100] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[101] Han Wang, Xiaojun Zhou, Qinyu Xu, Huaqiang Ren, Rong Xie, and Li Song. A large-scale sports tracking dataset and progressive re-detection based sports tracking. In *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2022.

[102] Zhaobin Wang and Zekun Yang. Review on image-stitching techniques. *Multimedia Systems*, 26(4):413–430, 2020.

[103] Qiang Yao, Keisuke Nonaka, Hiroshi Sankoh, and Sei Naito. Robust moving camera calibration for synthesizing free viewpoint soccer video. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1185–1189, 2016.

[104] Neng Zhang and Ebroul Izquierdo. A high accuracy camera calibration method for sport videos. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021.

[105] Neng Zhang and Ebroul Izquierdo. A fast and effective framework for camera calibration in sport videos. In *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2022.

Appendix A

Panoramas

A.1 Video 1 Results



Figure A.1: Panorama Video 1 - Manual Motion



Figure A.2: Panorama Video 1 - Optical Flow



Figure A.3: Panorama Video 1 - MPEG

A.2 Video 2 Results



Figure A.4: Panorama Video 2 - Optical Flow



Figure A.5: Panorama Video 2 - MPEG

A.3 Video 4 Results

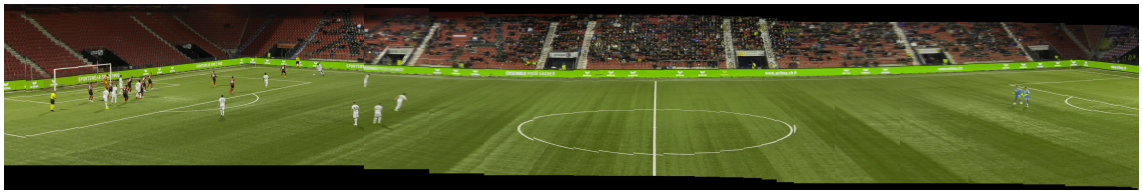


Figure A.6: Panorama Video 4 - Optical Flow

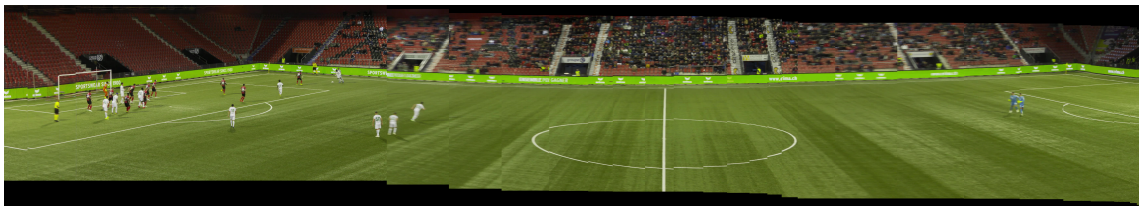


Figure A.7: Panorama Video 4 - MPEG

A.4 Video 5 Results



Figure A.8: Panorama Video 5 - Optical Flow



Figure A.9: Panorama Video 5 - MPEG