
# Occupancy Analysis of the Outdoor Football Fields

Huda, Noor Ul

*DOI (link to publication from Publisher):*
10.54337/aau561814357

*Publication date:*
2023

*Document Version*
Publisher's PDF, also known as Version of record


# OCCUPANCY ANALYSIS OF THE OUTDOOR FOOTBALL FIELDS

**BY**
**NOOR UL HUDA**

DISSERTATION SUBMITTED 2023

AALBORG UNIVERSITY
DENMARK

# Occupancy Analysis of the Outdoor Football Fields

Ph.D. Dissertation

Noor Ul Huda

Aalborg University
Department of Architecture, Design and Media Technology
Rendsburggade 14
DK-9000 Aalborg
Dissertation submitted month 06, 2023

Dissertation submitted: June 13, 2021

PhD supervisor: Assoc. Prof. Rikke Gade
Aalborg University

Assistant PhD supervisor: Prof. Thomas B. Moeslund
Aalborg University

PhD committee: Associate Professor Claus B. Madsen (chair)
Aalborg University, Denmark

Professor Dan Witzner Hansen
IT University of Copenhagen, Denmark

Associate Professor Toon Goedemé
Technical University De Nayer at
Sint-Katelijne-Waver, Belgium

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Architecture, Design
and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-691-1

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk



Printed in Denmark by Stibo Complete, 2023

# Curriculum Vitae

## Noor Ul Huda

Noor Ul Huda received her MSc in Electrical Engineering with a thesis based on medical image processing from the Center of advance studies in engineering (case) in Islamabad, Pakistan, in 2014. Before starting her ph.D. studies at Aalborg University, Denmark. She served as a lecturer at COMSATS University Islamabad, Pakistan, for almost 2.5 years. She started her PhD studies in Visual Analysis and Preception (VAP) lab at Aalborg University Denmark in April 2017.

During her Master's thesis, she worked on retinal image datasets for detecting neurovascular neuropathy for two well-reputed hospitals in Pakistan. She has also worked on some image- and video-based projects at COMSATS University. She was also involved in teaching and lab work monitoring for undergraduate students. Her area of interest is Image processing, computer vision and video data analysis.

Curriculum Vitae

# Abstract

This is the era of cost-effectiveness and resource management. The biggest and foremost concern of the managerial authorities is cutting the cost short while maximizing the benefits. Occupancy analysis is the solution to one such problem. The analysis is performed to monitor a place for a long time to figure out the time slots in which the particular place is not in use. This, in turn, can help the authorities utilize the space more efficiently.

This thesis aims to devise an effective and efficient method for occupancy analysis of outdoor football fields. Football/soccer is the most played sport in all over the world. It is well played across Europe regardless of the weather and environmental conditions. However, installing artificial grass in every other football field is expensive for the managerial authorities in harsh winter and sandy grounds like Denmark. Occupancy analysis of such fields can minimize this cost by figuring out the utilization of the fields vs the time of the day. In the long run, this analysis helps to figure out the fields or parts of the fields not used by the people at a particular time or day. This thesis covers the camera-based methods for the occupancy analysis of outdoor soccer fields.

Weather and light play a key role in selecting any camera setup to perform any vision-based analysis in an outdoor environment. Therefore, as the first and foremost study, preliminary research is conducted to explore different camera options to cover the whole soccer field in changing weather and light conditions. The analysis is performed by considering the cost-effectiveness and setup management complexity.

In the later section of this thesis, methods for player detection and occupancy analysis are presented for each type of camera setup. First, a method of combining distortion correction and appearance-based features for classical machine learning is introduced using a fisheye camera. The method is developed for player detection based on occupancy analysis of soccer fields.

Afterwards, using thermal cameras, two approaches for player detection are introduced. One uses machine learning and virtual reality-based features, and the other uses a deep network. The latter approach also discusses the role of different kinds of data in learning the convolution neural network

for person detection in thermal sensors and introduces a diverse thermal dataset for person detection. A study to observe the network behaviour by introducing homogeneity in data polarity for person representation is also conducted for a thermal camera.

Finally, this thesis introduces a method for monitoring the soccer field using one thermal and one fisheye camera. The thermal camera partially captures the soccer field in this setup and is fully captured by a fisheye camera. Furthermore, the method learns the person representations of missing thermal view using an adaptive student-teacher-based network. The final setup and the occupancy report in different fields of Aalborg, Denmark, is then provided to the municipality to enable them to figure out the ways of resource management and cost-effectiveness.

The thesis also leads to the publication of two new datasets for person detection, one for diverse outdoor conditions in outdoor thermal and another for cross-learning or multimodal and multiview distillation using thermal and fisheye camera feeds.

# Resumé

Omkostningseffektiv ressourcehåndtering er et væsentligt område for mange myndigheder. Her gælder det om at reducere omkostningerne mens fordelene maksimeres. For at opnå dette kan belægningsanalyser bruges som beslutningsstøtte til myndighederne. En belægningsanalyse udføres ved at overvåge et sted i lang tid for at finde ud af, hvilke tidspunkter stedet ikke er i brug. Dette kan hjælpe myndighederne med at udnytte pladsen mere effektivt.

Denne afhandling har til formål at udvikle en effektiv metode til belægningsanalyse af udendørs fodboldbaner. Fodbold er den mest spillede sport i hele verden og spilles på tværs af hele Europa uanset vejr- og miljøforholdende. I nogen lande som fx Danmark kan vejrforholdende medføre at almindelige fodboldbaner ikke kan bruges om vinteren. Ligeledes kan det være svært at lave almindelige fodboldbaner i sandede områder. Derfor anvendes der ofte kunstgræsbaner, men da kunstgræsbaner er dyre at lave er det vigtigt at banerne udnyttes fuldt ud. Belægningsanalyse af sådanne baner kan bruges til at undersøge hvornår og i hvilket omfang banerne anvendes. Disse informationer kan bruges som beslutningsgrundlag for hvorvidt der skal investeres i flere baner.

Denne afhandling omhandler kamerabasede metoder til belægningsanalyse af udendørs fodboldbaner. Når der skal laves vision-baserede analyser kræver det at det som skal analyseres fremstår tilstrækkeligt tydeligt på billederne fra kameraerne. I et udendørs miljø påvirkes billederne af vejr og lysforholdene og det er derfor nødvendigt at finde en kameraløsning som kan levere tilstrækkelig god billedkvalitet på alle tider af døgnet uanset tidligere nævnte forhold. Derfor blev der i det første studie undersøgt en række forskellige kameratyper og -opsætninger. Målet var at finde det mest omkostningseffektive setup der kunne dække en hel bane og levere billeder af tilfredsstillende kvalitet, uanset vejr og lysforhold.

I den efterfølgende sektion af afhandlingen præsenteres forskellige metoder til detektering af spillere og belægningsanalyser for hver kameraopsætning. Først introduceres en metode som kombinerer korrektion af forvrængning fra linsen og feature-baserede funktioner til klassisk maskinlæring ved an-

vendelse af et fiskeøjekamera. Derefter introduceres to metoder til spillerdetektering ved brug af termiske kameraer. Den ene metode bruger maskinlæring og virtual reality-baserede features, mens den anden bruger deep learning. I forbindelse med sidstnævnte metode diskuteres også betydningen af forskellige typer data ved træning af et convolutional neural network. Derudover præsenteres et termisk datasæt til persondetektering som indeholder termiske billeder for en række forskellige scenarier. En undersøgelse af netværkets opførsel ved introduktion af homogenitet in polariteten af de termiske billeder er ligeledes gennemført. Til sidst præsenterer denne afhandling en metode til overvågning af fodboldbanen med brug af et termisk og et fiskeøje kamera. Fiskeøjekameraet dækker hele banen mens det termiske kamera ikke dækker hele banen. Desuden lærer metoden vha. et student-teacher netværk hvordan en person er repræsenteret i de områder som ikke er dækket af det termiske kamera. Metoderne udviklet i dette projekt er blevet anvendt på forskellige baner i Aalborg i Danmark og er blevet præsenteret for kommunen som beslutningsstøtte i forbindelse med prioritering af kommunens ressourcer.

Afhandlingen har også ført til udgivelsen af to nye datasæt til persondetektering, et der indeholder termiske billeder til persondetektering under en række forskellige udendørs forhold, og et til tvær-læring eller multimodal og multivisual distillation ved brug af termisk og fiskeøje kamerafeed.

# Publications

Parts of the thesis work have been published in peer-reviewed scientific journals and international conferences.

## Journal Papers

1. Noor Ul Huda, Bolette Dybkjær Hansen, Rikke Gade, and Thomas B. Moeslund. "The effect of a diverse dataset for transfer learning in thermal person detection." In the Sensors, 20(7):1, April 2020. doi: 0.3390/s20071982.

## Conference Papers

1. Noor Ul Huda, Bolette D. Hansen, Rikke Gade, and Thomas B. Moeslund. "Occupancy analysis of soccer fields using wide-angle lens." In *The International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, pages 354–359, 9 apr. 2018. doi: 10.1109/SITIS.2017.65.

2. Noor Ul Huda, Kasper H. Jensen, Rikke Gade, and Thomas B. Moeslund. "Estimating the number of soccer players using simulation-based occlusion handling." In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1905–190509, June 2018. doi: 10.1109/CVPRW.2018.00236.

3. Anthony Cioppa, Adrien Deliège, Noor Ul Huda, Rikke Gade, Marc Van Droogenbroeck, and Thomas B. Moeslund. Multimodal and multi-view distillation for real-time player detection on a football field. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 880-881, June 2020. doi: 10.1109/CVPRW50498.2020.00448.

4. Noor Ul Huda, Rikke Gade, and Thomas B. Moeslund. "Effects of Preprocessing on the Performance of Transfer Learning Based Person Detection in Thermal Images,In *IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, pages. 86-91, August 2021. doi: 10.1109/PRML52754.2021.9520729.

# Contents

# Preface

This thesis is submitted to the Faculty of IT and Design, Aalborg University (AAU), with the fulfilment of the requirements for the Doctor of Philosophy. The work was carried out from April 2017 to February 2023 at the Department of Architecture, Design and Media Technology, AAU. The thesis work is supervised by Prof. Thomas B. Moeslund and co-supervised by Associate Prof. Rikke Gade.

I want to present my deepest regards and gratitude to both supervisors for providing me with the best support and assistance. Rikke was always very positive, with a smiling face and an encouraging attitude. She always gave me her time whenever I approached her. She listened to me in challenging situations and provided me with appropriate help. Thomas always pushed me in the right direction. He always kept his door open for discussion about any problem. The positive vibes and understanding from both of my supervisors and their support in many difficult times helped me achieve my goals and complete this thesis. I have never heard of any supervisor so supportive and considerate. They accommodated me in the best possible ways whenever I suffered mentally and physically.

I thank my colleagues Bolette, Anne and Chris for their help and friendship. It would have been challenging to stay in Denmark If they were not there with cheerful faces.

I want to express my gratitude to my father and my mother. I would not be there if it were not for them and because of them. Their support throughout my life enables me to explore more and more. They did their best to provide me with the best education. They gave me wings and freedom in an environment where most women are born only to do house chores.

Finally, this thesis is finished because of the love and support of my husband. If my parents gave me wings, he showed me places to fly with those wings. I would have left everything in the middle if he had not been there for me. He was always there to hear me out crying, stopping me from suicide, counselling me, and even writing my dictations during my days of sickness. Therefore, I dedicate my thesis to the most supportive, caring, and giving person.

Thank You, Allah Talaa, for filling my life with the most supportive and loving people. Thank you for blessing me with the best among your people.

Noor Ul Huda
Aalborg University, June 13, 2023

# Part I

# Introduction

# Chapter 1

# Introduction

The thesis aims to develop a robust algorithm to perform long-term occupancy analysis in outdoor soccer fields in Denmark. In this work, different camera setups are proposed and analyzed for person/player detection for finding occupancy in the local outdoor soccer fields.

## 1 Motivation and background

Soccer is the most watched and played sport in the world [1]. People all over the world follow and play both indoor and outdoor soccer. Costly and complex camera setups are utilized for the commercial country club and international matches. These camera setups not only record but also focus on different game angles to use the information for game analysis, prediction, players' performance monitoring, predicting trends, reviewing decisions in video-assisted referees and many more. However, when it comes to local soccer, the primary concern of the management is cost minimization.

The weather conditions in the North of the world are harsh, especially in winter. It is almost impossible to maintain good ground conditions on soccer fields. The optimum solution to maintain a winter soccer field is to install artificial grass, which is expensive. The management always looks for optimum solutions to such problems. In such a scenario, occupancy analysis of the targeted fields can provide the required information about their usage in order to help management optimize the expenses. Because there may be many fields around that are occupied only some of the time. In some scenarios, people book the time but do not show up at the field or may be so small in numbers that they could use half of the field. The resources in such scenarios can be optimized. An example of how an occupancy analysis chart looks is illustrated in Fig. 1.1.


**Fig. 1.1:** Example of occupancy chart for seven days(monday to saturday, from 7:00 am to 11:pm). The grey rows in the chart are actual bookings, while the coloured rows show the occupancy. The colours from green to red illustrate minimum to maximum occupancy at respective times.

The traditional method for occupancy analysis involves a human annotator, which is expensive and inconvenient for long-term projects. Now with the emergence of vision-based technologies, imaging sensors are widely employed almost everywhere. Therefore, vision-based occupancy analysis methods can be proposed to save time and cost.

With the given considerations, this thesis follows the motivation to provide the municipality with a robust solution for occupancy analysis for outdoor soccer fields using a vision-based setup. Therefore, the thesis revolves around the application challenges for different camera setups and ends with the final report to the municipality.

The main challenge in the vision-based method lies in understanding the outdoor weather conditions and the choice of image sensor that has good performance for a long time. In the context of weather conditions, the problem with the outdoor data lies in its uncertainty due to the presence of noise factors, including wind, varying sunlight, night vision, rain, snow, shadows and many more. Another challenge is to keep the overall system low budget and easy to install, as the fields are armature and mostly need to be equipped with proper lighting and installation setups (like mounting pole, electric and network supply).

# 2 State of the art

## 2.1 Computer vision in sports

For decades camera systems have been widely implemented in different sports areas. The application varies from broadcasting to player performance

analysis [2]. Many camera-based systems are used commercially [3–9] as well as in local sports arenas for a wide range of applications [10–16, 16–20]. Thomas et al. [2] discuss in detail the current applications of vision in sports. The article also describes the commercially used setups for monitoring and broadcasting in sports. It also provides the details about many publically available datasets for vision-based applications like player detection [21–23], ball detection [21], tracking [21, 24, 25], camera calibration [26], event detection [27, 28] and many more.

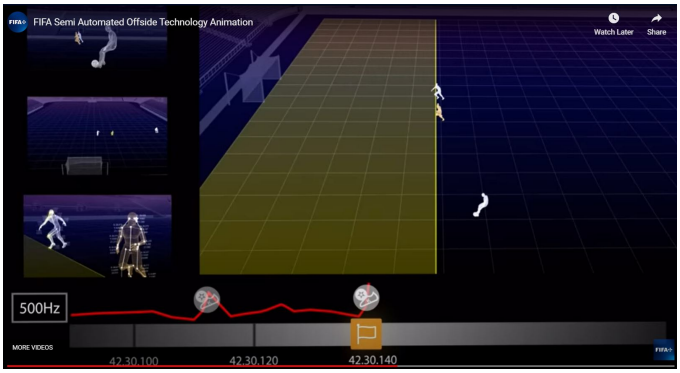



**Fig. 1.2:** example of computer vision in Sports. Semi-automated offside technology implemented in FIFA 2022 [29].

The primary application domains for applicable computer vision in soccer or football analysis are mainly match analysis [10, 30–35], player detection [25, 36–38], action recognition [39], event detection and recognition [14] object tracking [15, 40, 41], game performance analysis [16, 42], camera setups [26], scene analysis [18], occupancy analysis of fields [19], audience/ crowd Analysis [20], and game analysis [16, 42].

Commercially available solutions mainly focus on video base analysis by manually tracking the ball and players or providing GPS-based support. SPIIDEO provides the solution for performance analysis and camera systems for proper coverage [3]. Hudl sportscode provides tools for performance analysis [4]. Inmotio [8] provides solutions like player tracking, action generation and data analysis. iSports analysis, Nacsports, Interplay [5–7] provides packages to players and coaches for performance monitoring.

Research has been conducted in action detection, activity detection and action recognition in sports using vision-based methods. Vanderplaetse et al. performed action spotting using visual and audio cues at different stages of DNN [43]. Kanimozhi et al. proposed a content-based viewpoint for classifying meaningful events in sports videos [44]. Host et al. tested the baseline CNN model with LTSM and MLP-based models for classifying 11 actions in

handball matches [45]. Piergiovanni et al. introduced MLB-Youtube data set for activity detection and tested a range of algorithms on video streams of baseball broadcast [46]. Sanford et al. model the ball and player interaction events and used self-attention in soccer play for group activity detection [28]. Schlosser et al. tested two-stream architectures to generate proposals for temporal actions [47]. Rahimi et al. presented a technique based on extracting scene graph features from sports videos [48]. Giancola et al. proposed a feature pooling-based method on NetVLAD, dubbed NetVLAD++. Instead of creating a single pool, the method splits the context before and after the occurrence of an action. They claimed that including both the prior and posterior information creates a more distinctive pool of features for understanding actions [49]. Hong et al. proposed a method based on pose estimation for action recognition in sports videos. They introduced a video pose distillation method to learn features from the video domain in a student-teacher manner [39]. Many others [50–52] also utilized CNN-based approaches to action recognition in sports.

In sports, professionals are also eagerly interested in the occurrence of meaningful events, i.e. opportunities created, attacks and others, to assess teams or individual performances and make plans for coming games. To facilitate such an approach, Cioppa et al. [53] have proposed a two-step process for event detection. In the first step, a deep learning-based network is built to extract semantic features, which in the second step are passed to a decision tree classifier for event detection in the soccer game. Kanojia et al. [54] proposed an LSTM-based model to identify and classify 48 events/tasks that are performed in diving. Vats et al. [55] developed a model for event detection for the ice hockey game. The method combined a single 2D CNN for features extraction and multiple 1D CNN architectures with varying kernel sizes to detect actions for identifying events. Kaichi et al. [27] investigated the application of camera vision for analyzing athletes' performance by determining the centre of mass (CoM). Shukla et al. [56] worked on automatic highlight generation for cricket matches. Saikat et al. [57] employed SVM based approach to map the game scenarios and generate game statistics. Shukla et al. [56] worked on automatic highlight generation for cricket matches. Saikat et al. [57] employed SVM based approach to map the game scenarios and generate game statistics.

Cioppa et al. [26], Kosuke et al. [58], Jianhui et al. [17] worked on camera calibration methods in sports video recordings. A method that is the combination of segmentation, encoding of zone segmentation and template homography to obtain calibration parameters is implemented by Cioppa et al. [26]. Kosuke et al. [58] proposed an alternative method of processing images independent of calibration and synchronization to estimate 2D and 3D human poses. In another approach, Jianhui et al. employed an automated process using synthetic data and a generative adversarial network (GAN)

model for camera calibration [17].

Many new datasets have also been published in sports for research purposes. For example, Giancola et al. [59] have collected and made the dataset of soccer matches public. The dataset comprises 500 games with a total duration of 764 hours. The dataset has already been annotated for the goals, yellow/red cards and substitution events. Frame rates of these videos vary between 25 and 50 fps with MPEG and H264 encoding and SD/full resolution. Tanaka et al. [60] prepared the dataset from the game "League of Legends". The authors collected a total of 9723 clips and 62677 captions from the videos of the game played at the world championship. Deliege et al. [61] published the annotations of 300k images within soccerNet for action spotting, segmentation and boundary detection.

# 3 Player detection

Player detection is the primary challenge in many of the vision-based sports applications, i.e. tracking [12], action recognition [62], and player recognition [63]. It has been performed for both outdoor and indoor sports, including soccer [11], badminton [64, 65], baseball [63, 66], basketball [11, 25, 67–69], hockey [70, 71], handball [72, 73], tennis [12, 36, 74], squash [75] and running [76].

Several techniques have been reported for player detection. Of these techniques, background subtraction [77] is the most frequently used method for this purpose. It has been reported for its fast-processing time, which is favourable for real-time applications. Moreover, It can be applied for both static [36, 37] and moving cameras if the surfaces have uniformity [78]. Other algorithms, including Otsu, edge-based detection, haar-like features, and hough transform [79–81] have also been implemented for player detection in the soccer field.

Background uniformity in outdoor sports is of significant concern for various reasons, e.g. colour contrast between background and foreground, varying lighting conditions, and many others. Several methods have been proposed to address these challenges. SVM has been widely used because of its low computation complexity, optimal solution and better accuracy in detecting unseen data [82, 83]. In [82], Zhu et al. implemented SVM for player detection using a player colour model defined in HSV colour space. A particle filter-based SVR algorithm is implemented for player tracking, followed by a player detection algorithm. Maćkowiak et al. [83] also used SVM for player detection in a broadcast soccer video. In their work, a colour segmentation-based playfield detection algorithm is also implemented—afterwards, the HOG features-based SVM classifier is used for player detection. Bai et al. [84] proposed a method for automatically generating labelled data. They imple-

mented a one-class SVM technique to detect players on a soccer field. One-class SVM has also been reported in [85], where it is followed by a fuzzy c-mean algorithm to facilitate the prediction of data points close to the SVM hyper-plane. In [38], HOG features-based Adaboost algorithm is proposed for player detection on the football field. An unsupervised approach using contrastive learning from Jersey colours is proposed by Koshkina et al. [86]. Pidaparthy et al. [87] used deep visual and audio cues to identify the active period in the game to aid player segmentation in hockey using the hidden Markov model.

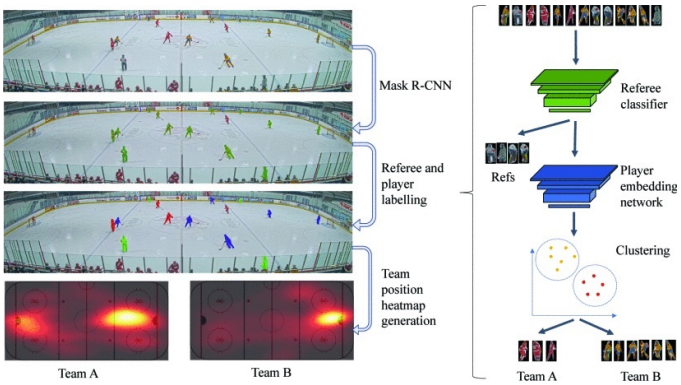



**Fig. 1.3:** Overview of the proposed system. Mask R-CNN is first used to detect and segment each person on the playing surface. A pre-trained CNN is then used to classify referees while the remaining players are passed to our embedding network for clustering into teams. This allows the production of heat maps showing the distribution of the two teams over the playing surface. [86]

With the recent advancement, DNNs being the standard for solving many object detection problems, YOLO [21–25, 88, 89], SSD [88] and Faster RCNN [72, 88, 90–92], UNet [93], and many other deep networks have been utilized for player detection in static and moving cameras. In [94], shallow and reverse-connected CNN networks are proposed for player detection. Sah et al. also compare deep learning methods with traditional methods for player detection in field hockey datasets [88]. Previously they compared different image representations to feed in one CNN for the same hockey dataset for player detection [94]. Zhang et al. [21], Kalafatic et al. [24], You et al. [95], Acuna et al. [25], Pobar et al. [22], Buric et al. [23, 89] implemented YOLO Network for the purpose of player detection in sports videos. Zhang et al. [21] combined YOLO V4 [96] with Deep Sort [97] to detect and track players and football in NBA and Worldcup matches. Kalafatic et al. utilized YOLO V3 [98] with Faster R-CNN anchor boxes for player detection and SORT algorithm for player tracking in football fields. They compare their work with traditional adaptive background subtraction-based player detection systems.

Buric et al. [23, 89, 99] combine YOLO with optical flow for player detection and tracking.

Acuna et al. [25] also used YOLO with SORT [100] algorithm to detect and track football players. They used YOLO V2 [101] for the purpose of detection. A method based on adding reverse-connected modules to CNN for multiscale player detection is presented by Zhang et al. [102]. Theagarajan et al. [103] proposed DCGAN for data augmentation and found it to help improve the object detection accuracy inside a soccer field. Lu et al. [104] implemented a cascaded CNN to detect players in varying light conditions, low-quality images and high-speed moving cameras. A method based on the fisher vector combined with CNN to identify players in basketball is presented by Senocak et al. [105]. Komorowski et al. designed and implemented the pyramid-based deep neural network Deepball [106] for ball detection and dubbed FootAndBall [91], designed for player and ball detection in high-resolution video recordings.

## 3.1 Occupancy analysis

Occupancy analysis is an essential element for managing resources effectively and efficiently. In literature, methods to perform occupancy analysis can be divided into two categories, i.e. visual or non-visual. In non-visual approaches, multiple sensor data is used, and most proposed solutions are targeted for indoor occupancy analysis.

Zhao et al. [107] utilized the data of multiple sensor networks, i.e. WiFi, GPS and chair/keyboard/mouse sensors, to implement a Bayesian belief network and perform occupancy analysis for offices. Other methods proposed using the data from passive infrared sensor [108–110] and Carbon dioxide (CO2) level [111, 112]. Dong et al. [113] developed a test bed comprised of wireless and wired sensor networks, i.e. wireless ambient sensing system, CO2 and air quality reader sensors, to analyze occupancy. They implemented support vector machines (SVM), neural networks (NN) and hidden Markov models (HMM) techniques to process sensor data. Pedersen et al. [114] also proposed climate sensor data, i.e. CO2, PIR and volatile organic compound, to detect room occupancy. Xin et al. [115] proposed a data mining approach to perform occupancy analysis by only using the time series data of people inside the building. The aforementioned non-visual methods have the disadvantage of a limited application area, as they are only feasible for buildings and closed rooms but not for large outdoor fields.

To overcome the challenges of occupancy analysis in an outdoor environment, image-based techniques are proposed [116–118]. The methods of processing images or video sequences for person detection, where the data is recorded either using RGB or thermal cameras. The applications of these methods are found in indoor offices, pedestrian detection and players detec-

tion in sports fields [114, 118–120]. Specifically, for the sports arena, little literature has been published in the domain of thermal cameras. Gade et al. [119], and [120] performed occupancy analysis for indoor sports arenas using thermal data. Occupancy analyses for outdoors have also been reported for parking lots [121–123].

## 3.2 General Challanges

Within the image-based techniques, many challenges exist when applied to outdoor environments, i.e. occlusion, weather/light conditions and monitoring large areas. A brief state of the art in these domains is discussed below.

- **Occlusion:** Occlusion is a primary challenge in performing long-term occupancy analysis and many other vision analysis applications. In this phenomenon, persons very close to each other can be identified as one, affecting the accuracy and reliability of long-term occupancy and people detection results. Some examples are shown in Fig. 1.4

**(a)** **(b)** **(c)**

**Fig. 1.4:** Examples of occlusion in **(a)** Thermal camera, **(b)** RGB camera and **(c)** Wide-angle RGB camera.

Many methods [124–129] have been proposed to meet this challenge. In [130], Jin et al. presented a human tracking algorithm using an RGB camera. In the proposed method, the occlusion is handled by applying a threshold to the output of the target person template and given frame comparison. Marin et al. [131] handled the partial occlusion for human detection application. The reported method is comprised of two steps. At first, a holistic classifier is implemented, whose output is further processed by an ensemble classifier for human detection if the confidence output of the first classifier lies in a defined range referring to possible occlusion. In [132], Zhou et al. a bi-box regression approach to estimate the occlusion and simultaneously detect a pedestrian. In the proposed method, two branch CNN is implemented, one targeted for regressing the bounding box for the entire body and the other for visible human parts. In [133], an RNN algorithm is proposed to detect occlusion while multiple persons. In [134], Shu et al. developed a part-base model to detect partial occlusion in multi-person tracking prob-

lems. Solutions to deal with partial occlusion for detection, classification and tracking applications have also been reported in [135–137]. In these works, the Bhattacharyya distance, pyramidal part-based model and restricted Boltzmann machine-based deep models are developed to detect occlusion.

- **Weather effects:** Weather effects that include the sun, shadow, rain, and surrounding and body temperatures affect the output of cameras used and, in turn, the performance of person detection and occupancy analysis.

(a) (b)

**Fig. 1.5:** Examples of the effect of sunny weather in **(a)** RGB camera, **(b)** Thermal camera.

The shadows affect the performance of the detection system by distorting the person representation [138]. A four-stage shadow detection and classification approach is presented in [139]. In the proposed algorithm, the candidate region is passed through a weak classifier to detect shadow points, followed by determining spatial and temporal constancy. Finally, the comparison of the last two stages gives the shadow classification. Methods for shadow detection are also being studied in [140]. Kristo et al. [141] studied the object detection performance of thermal data in the presence of rain, fog and precise weather conditions. They implemented YOLOv3 and compared its performance with three other algorithms, including Faster R-CNN, SSD, and Cascade R-CNN. Tumas et al. [142] also studied thermal data in varying weather conditions, i.e. clear, cloudy, rain and fog. They also used YOLOv3 deep network to deal with these problems.

- **Large area coverage:** In the domain of occupancy analysis, especially in outdoor coverage of an extensive area is a primary challenge. Moreover, issues can arise in terms of cost, maintenance complexity and performance while selecting cameras. Thus, the camera selection becomes challenging to balance all these measures.

  In the RGB domain, different cameras are available depending on the field of view. Wide angle cameras have the advantage of wide viewing angle, and for this reason, it has been reported for application, i.e. surveillance [143, 144], indoor environment [145], automobiles [146].

Deep learning-based approaches are getting popularity in object detection in recent times. Pre-trained deep networks and transfer learning are other possibilities. In the last decade, many deep learning-based networks [98, 147–152] have been abundantly created and utilized for person detection in colour images.

Thermal cameras can perform very well at night and in low-light conditions. Moreover, the results are not affected by foreground and background colour similarity. Thermal soccer dataset [119], OSU thermal pedestrian [153], OSU Color Thermal [154], Terravic Motion IR [155], and CVC-09 [156] is the publically available thermal datasets. These datasets have been recorded for pedestrian detection, person detection in the outdoor and indoor environment and player detection in the sports arena. traditional methods based on feature extraction, and thresholding [118–120, 157] have been enormously employed for person detection in thermal. Machine learning [158, 159] combined with Histogram of gradients and deep neural networks [160, 161] have also been used in recent studies for person detection in thermal cameras.

If we look for the comprehensive dataset for local/non-perfect labelled data, there is a bridge/gap. It shows that the available datasets and the designed algorithms are applied for ideal, predictable conditions obtained from closed environmental conditions. Moreover, different camera setups can behave differently in given conditions for a particular application in an outdoor environment. Therefore, this thesis analyses the challenges mentioned above by employing and testing multiple camera setups for long-term outdoor monitoring.

Moreover, The work presented in this thesis is inspired by [119] and [120], which was aimed at an indoor sports arena, whereas, here, we focus on an occupancy analysis system for the outdoor sports field.

# 4 Objectives and scope of the work

This PhD work aims to develop a robust algorithm for counting the number of persons in the soccer fields to identify the occupancy measure for a very long time. Therefore, in this thesis, different camera setups will be presented. Furthermore, the person detection methods for different camera setups will be developed and tested to determine the desired occupancy in an unstable outdoor environment. To this end, the following research tasks will be conducted:

- Investigate feasibility analysis of different camera setups for occupancy analysis for large area coverage while keeping the system simple and cost-effective.

- Develop an algorithm for the robust occupancy analysis of outdoor fields that performs well even in challenging weather, wind and light conditions. The algorithm should be able to identify and classify the occlusions to count the number of players for occupancy.
- Evaluate algorithm performance on raw data from the fields in Aalborg, Denmark.
- Testing and performing long-term occupancy analysis in the soccer fields of Aalborg, Denmark and reporting to the municipalities.

Fig. 1.6 shows the overall scope of the thesis along with the addressed challenges and related thesis chapters. The following section describes the overall structure of the thesis.

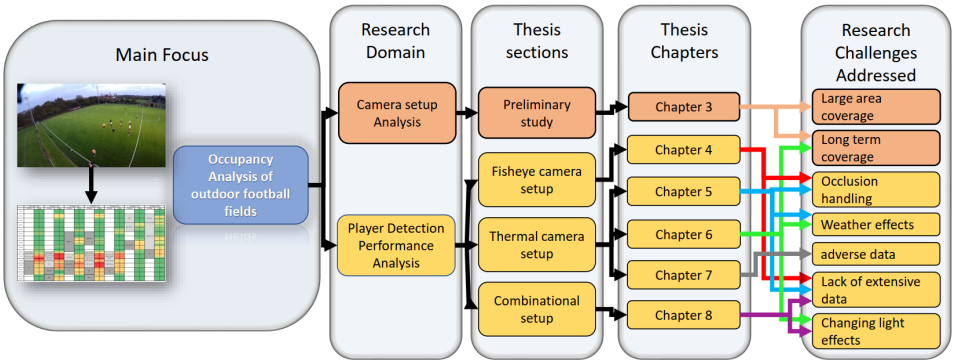



**Fig. 1.6:** scope of the thesis.

# 5 Outline of the thesis

The thesis is organized into six sections indicated by roman numbers. The first section provides an introduction to the thesis, including motivation and a summary of the conducted research. The following four sections relate to the sensor approach, see Fig. 1.6, and the last section concludes the thesis. Each section consists of one or more chapters. The chapters of sections II-V are indicated in Fig. 1.6. Chapters 9 and 10 belong to part VI. Below is a summary of the different chapters.

Chapter 1 explains the background of the research, motivation, scope of the thesis and state of the art in the field. Chapter 2 provides a summary of the coming chapters.

Chapter 3 of the thesis is based on the feasibility study of different camera setups for outdoor long-term occupancy analysis. Analyzing setups analyt-

ically is the first and foremost task. Therefore, the chapter first defines and evaluates different camera setups based on observation and data analysis.

An algorithm for player detection and finding occupancy in local outdoor soccer fields using a fisheye camera setup is presented in chapter 4. The challenges in the fisheye camera image outdoors include very high distortion at the corners and occlusion in players that worsen in the presence of high camera distortion.

Chapters 5, chapter 6 and chapter 7 deal with some of the challenges in thermal camera setup. The challenges include occlusion detection and handling, weather challenges and lack of enough thermal data to deal with such challenges when implementing deep learning. Chapter 5 presents an algorithm for detecting players in the field and handling occlusion using machine learning and virtually projected features by 3D simulations. Chapter 6 reviews the thermal person detection datasets and presents a diverse dataset that includes varying weather and light effects. The chapter also studies person detection in thermal images using transfer learning from RGB and indoor thermal data. Finally, chapter 7 explores preprocessing methods' effect on creating homogeneity in data polarity in thermal person detection datasets. The study defines different preprocessing methods and investigates the outcomes using transfer learning and the diverse dataset published in chapter 6.

Chapter 8 presents a method based on multi-modal and multi-view distillation. It deals with the challenge of changing light conditions day and night by presenting a relatively cheap and straightforward installation setup of one thermal to only see in the middle of the field and one fisheye camera to cover the whole field.

Chapter 9 discusses the final setup for long-term occupancy, the report and results to the municipality and chapter 10 conclude the research with possible future directions.

The research conducted in this thesis is carried out by considering the applicability of the algorithms in practical scenarios. All the experiments are performed on self-recorded data from actual outdoor soccer fields. Thus, it covers all the real scenarios and diversity w.r.t weather, light conditions, and person representations.

# References

[1] J. Shvili, "The most popular sports in the world," Oct 2020. [Online]. Available: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

[2] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton, "Computer vision for sports: Current applications and research topics," *Computer Vision and Image Understanding*, vol. 159, pp. 3–18, 2017, computer Vision in Sports.

[3] "Spiideo," July 2021. [Online]. Available: https://resources.spiideo.com/videoanalysis-uk-football-lp?utm_term=football%20camera%20systems&utm_campaign=Resultify+%7C+Football+UK&utm_source=adwords&utm_medium=ppc&hsa_acc=9073649153&hsa_cam=1683470074&hsa_grp=65462553837&hsa_ad=522446653653&hsa_src=g&hsa_tgt=aud-644659769915:kwd-303642740593&hsa_kw=football%20camera%20systems&hsa_mt=b&hsa_net=adwords&hsa_ver=3&gclid=CjwKCAjwz_WGBhA1EiwAUAxIcSivM9yVGb_GJJskDt2xh94QFTcUbsauS3I7fj0jugtChdKxGsO0RRoCYyYQAvD_BwE

[4] H. Sportscode, Nov. 2022. [Online]. Available: https://www.hudl.com/products/sportscode

[5] Nacsport, Nov. 2022. [Online]. Available: https://www.nacsport.com/index.php?lc=en-gb

[6] isportsanalysis, Nov. 2022. [Online]. Available: https://www.isportsanalysis.com/index.php

[7] interplay sports, Nov. 2022. [Online]. Available: https://interplay-sports.com/about-interplay-sports/

[8] inmotio, Nov. 2022. [Online]. Available: https://inmotio.eu/

[9] "Stemmer imaging," July 2021. [Online]. Available: https://www.stemmer-imaging.com/en/applications/imaging-systems-for-sports-tracking/

[10] D. Memmert, B. Noël, D. Machlitt, J. van der Kamp, and M. Weigelt, "The role of different directions of attention on the extent of implicit perception in soccer penalty kicking," *Human Movement Science*, vol. 70, p. 102586, 2020.

[11] D. Delannay, N. Danhier, and C. De Vleeschouwer, "Detection and recognition of sports(wo)men from multiple views," in *International Conference on Distributed Smart Cameras (ICDSC)*. ACM/IEEE, 2009, pp. 1–7.

[12] Y. C. Jiang, K. T. Lai, C. H. Hsieh, and M. F. Lai, "Player detection and tracking in broadcast tennis video," in *Advances in Image and Video Technology*, T. Wada, F. Huang, and S. Lin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 759–770.

[13] A. Sen and K. Deb, "Categorization of actions in soccer videos using a combination of transfer learning and gated recurrent unit," *ICT Express*, 2021.

[14] L. Morra, F. Manigrasso, and F. Lamberti, "Soccer: Computer graphics meets sports analytics for soccer event recognition," *SoftwareX*, vol. 12, p. 100612, 2020.

[15] W. Kim, "Multiple object tracking in soccer videos using topographic surface analysis," *Journal of Visual Communication and Image Representation*, vol. 65, p. 102683, 2019.

[16] M. Marchiori and M. de Vecchi, "Secrets of soccer: Neural network flows and game performance," *Computers and Electrical Engineering*, vol. 81, p. 106505, 2020.

[17] J. Chen and J. J. Little, "Where should cameras look at soccer games: Improving smoothness using the overlapped hidden markov model," *Computer Vision and Image Understanding*, vol. 159, pp. 59–73, 2017, computer Vision in Sports.

[18] X. Gao, Z. Niu, D. Tao, and X. Li, "Non-goal scene analysis for soccer video," *Neurocomputing*, vol. 74, no. 4, pp. 540–548, 2011.

[19] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine Vision and Applications, Springer*, vol. 25, no. 1, pp. 145–262, 2014.

[20] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 12, pp. 796–807, 02 2003.

[21] Y. Zhang, Z. Chen, and B. Wei, "A sport athlete object tracking based on deep sort and yolo v4 in case of camera movement," in *International Conference on Computer and Communications (ICCC)*, 2020, pp. 1312–1316.

[22] M. Pobar and M. Ivasic-Kos, "Active player detection in handball scenes based on activity measures," *Sensors*, vol. 20, no. 5, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/5/1475

[23] M. Buric, M. Pobar, and M. Ivašić-Kos, "Object detection in sports videos," in *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 1034–1039.

[24] Z. Kalafatić, T. Hrkać, and K. Brkić, "Multiple object tracking for football game analysis," in *International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2022, pp. 936–941.

[25] D. Acuna, "Towards real-time detection and tracking of basketball players using deep neural networks," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[26] A. Cioppa, A. Deliege, F. Magera, S. Giancola, O. Barnich, B. Ghanem, and M. Van Droogenbroeck, "Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2021, pp. 4537–4546.

[27] T. Kaichi, S. Mori, H. Saito, K. Takahashi, D. Mikami, M. Isogawa, and H. Kimata, "Estimation of center of mass for sports scene using weighted visual hull," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2018.

[28] R. Sanford, S. Gorji, L. G. Hafemann, B. Pourbabaee, and M. Javan, "Group activity detection from trajectory and video data in soccer," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2020.

[29] "Semi-automated offside technology," *FIFA*. [Online]. Available: https://www.fifa.com/technical/football-technology/

football-technologies-and-innovations-at-the-fifa-world-cup-2022/semi-automated-offside-technology

[30] V. Machado, R. Leite, F. Moura, S. Cunha, F. Sadlo, and J. L. Sadlo, "Visual soccer match analysis using spatiotemporal positions of players," *Computers and Graphics*, vol. 68, pp. 84–95, 2017.

[31] J. Gudmundsson and T. Wolle, "Football analysis using spatio-temporal tools," *Computers, Environment and Urban Systems*, vol. 47, pp. 16–27, 2014, progress in Movement Analysis – Experiences with Real Data.

[32] N. Vago, Y. Lavinas, D. Rodrigues, F. Moura, S. Cunha, C. d. C. Aranha, and R. Torres, "Integra: An open tool to support graph-based change pattern analyses in simulated football matches," 6 2020.

[33] Z. Niu, X. Gao, and Q. Tian, "Tactic analysis based on real-world ball trajectory in soccer video," *Pattern Recognition*, vol. 45, no. 5, pp. 1937–1947, 2012.

[34] M. Rafiq, G. Rafiq, R. Agyeman, G. S. Choi, and S.-I. Jin, "Scene classification for sports video summarization using transfer learning," *Sensors*, vol. 20, no. 6, 2020.

[35] E. Külah and H. Alemdar, "Quantifying the value of sprints in elite football using spatial cohesive networks," *Chaos, Solitons & Fractals*, vol. 139, p. 110306, 2020.

[36] M. Archana and M. Kalaiselvi Geetha, "An efficient ball and player detection in broadcast tennis video," in *Intelligent Systems Technologies and Applications*. Cham: Springer International Publishing, 2016, pp. 427–436.

[37] V. Renò, N. Mosca, M. Nitti, T. Dorazio, D. Campagnoli, A. Prati, and E. Stella, "Tennis player segmentation for semantic behavior analysis," in *International Conference on Computer Vision (ICCV) Workshop*. IEEE, Dec. 2015, pp. 718–725.

[38] H. Faulkner and A. Dick, "AFL player detection and tracking," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, Nov. 2015, pp. 1–8.

[39] J. Hong, M. Fisher, M. Gharbi, and K. Fatahalian, "Video pose distillation for few-shot, fine-grained sports action recognition," in *International Conference on Computer Vision (ICCV)*. IEEE/CVF, October 2021, pp. 9254–9263.

[40] X. Jiang, "Human tracking of track and field athletes based on fpga and computer vision," *Microprocessors and Microsystems*, vol. 83, p. 104020, 2021.

[41] C. G. Kagalagomb and S. Dixit, "Tracking of soccer players using optical flow," in *International Conference on Innovative Computing and Communications*, D. Gupta, A. Khanna, S. Bhattacharyya, A. E. Hassanien, S. Anand, and A. Jaiswal, Eds. Singapore: Springer Singapore, 2021, pp. 117–129.

[42] X. Meng, Z. Li, S. Wang, A. Karambakhsh, B. Sheng, P. Yang, P. Li, and L. Mao, "A video information driven football recommendation system," *Computers and Electrical Engineering*, vol. 85, p. 106699, 2020.

[43] B. Vanderplaetse and S. Dupont, "Improved soccer action spotting using both audio and video streams," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2020.

[44] S. Kanimozhi, T. Mala, A. Kaviya, M. Pavithra, and P. Vishali, "Key object classification for action recognition in tennis using cognitive mask rcnn," in *International Conference on Data Science and Applications*, M. Saraswat, S. Roy, C. Chowdhury, and A. H. Gandomi, Eds. Singapore: Springer Singapore, 2022, pp. 121–128.

[45] K. Host, M. Ivasic-Kos, and M. Pobar, "Action recognition in handball scenes," in *Intelligent Computing*, K. Arai, Ed. Cham: Springer International Publishing, 2022, pp. 645–656.

[46] A. Piergiovanni and M. S. Ryoo, "Fine-grained activity recognition in baseball videos," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2018.

[47] P. Schlosser, D. Munch, and M. Arens, "Investigation on combining 3d convolution of image data and optical flow to generate temporal action proposals," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2019.

[48] A. M. Rahimi, K. Lee, A. Agarwal, H. Kwon, and R. Bhattacharyya, "Toward improving the visual characterization of sport activities with abstracted scene graphs," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2021, pp. 4500–4507.

[49] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2018.

[50] T. Wang, C.-W. Chang, and Y.-S. Wu, "Template-based people detection using a single downward-viewing fisheye camera," in *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. IEEE, Nov. 2017, pp. 719–723.

[51] J. Liu and Y. Che, "Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network," *Journal of Electronic Imaging*, vol. 30, no. 3, pp. 1 – 16, 2021.

[52] M. Tabish, Z.-u.-R. Tanooli, and M. Shaheen, "Activity recognition framework in sports videos," *Multimedia Tools and Applications*, Feb 2021.

[53] A. Cioppa, A. Deliege, and M. Van Droogenbroeck, "A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2018.

[54] G. Kanojia, S. Kumawat, and S. Raman, "Attentive spatio-temporal representation learning for diving classification," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2019.

[55] K. Vats, M. Fani, P. Walters, D. A. Clausi, and J. Zelek, "Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2020.

[56] P. Shukla, H. Sadana, A. Bansal, D. Verma, C. Elmadjian, B. Raman, and M. Turk, "Automatic cricket highlight generation using event-driven and excitement-based features," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2018.

[57] S. Sarkar, A. Chakrabarti, and D. Prasad Mukherjee, "Generation of ball possession statistics in soccer using minimum-cost flow network," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2019.

[58] K. Takahashi, D. Mikami, M. Isogawa, and H. Kimata, "Human pose as calibration pattern; 3d human pose estimation with multiple unsynchronized and uncalibrated cameras," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2018.

[59] S. Giancola and B. Ghanem, "Temporally-aware feature pooling for action spotting in soccer broadcasts," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2021, pp. 4490–4499.

[60] T. Tanaka and E. Simo-Serra, "Lol-v2t: Large-scale esports video description dataset," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2021, pp. 4557–4566.

[61] A. Deliege, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck, "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2021, pp. 4508–4519.

[62] Q. Tran, B. Vo, T. Dinh, and D. Duong, "Automatic player detection, tracking and mapping to field model for broadcast soccer videos," in *International Conference on Advances in Mobile Computing and Multimedia*, ser. MoMM '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 240–243.

[63] Z. Mahmood, T. Ali, and S. Khattak, "Automatic player detection and recognition in images using adaboost," in *International Bhurban Conference on Applied Sciences Technology (IBCAST)*, 2012, pp. 64–69.

[64] Y. Peng, X. Ma, X. Gao, and F. Zhou, "Background estimation and player detection in badminton video clips using histogram of pixel values along temporal dimension," in *International Conference on Electronics and Information Engineering*, Q. Zhang, Ed., vol. 9794, International Society for Optics and Photonics. SPIE, 2015, p. 979409.

[65] N. Rahmad, N. A. J. Sufri, N. H. Muzamil, and M. A. As'ari, "Badminton player detection using faster region convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, pp. 1330–1335, 2019.

[66] Z. Mahmood, T. Ali, S. Khattak, L. Hasan, and S. Khan, "Automatic player detection and identification for sports entertainment applications," *Pattern Analysis and Applications*, vol. 18, 11 2015.

[67] Z. Ivankovic, M. Racković, and M. Ivkovic, "Automatic player position detection in basketball games," *Multimedia Tools and Applications*, vol. 72, 10 2014.

[68] B. Markoski, Z. Ivankovic, L. Ratgeber, P. Pecev, and D. Glušac, "Application of adaboost algorithm in basketball player detection," *Acta Polytechnica Hungarica*, vol. 12, pp. 189–207, 01 2015.

[69] M. Ivasic-Kos, M. Pobar, and J. Gonzàlez, "Active player detection in handball videos using optical flow and stips based measures," in *International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2019, pp. 1–8.

[70] C. Direkoglu, M. Sah, and N. E. O'connor, "Player detection in field sports," *Mach. Vision Appl.*, vol. 29, no. 2, p. 187–206, Feb. 2018. [Online]. Available: https://doi.org/10.1007/s00138-017-0893-8

[71] T. Guo, K. Tao, Q. Hu, and Y. Shen, "Detection of ice hockey players and teams via a two-phase cascaded cnn model," *IEEE Access*, vol. 8, pp. 195 062–195 073, 2020.

[72] M. Pobar and M. I.-K. Ivašić-Kos, "Detection of the leading player in handball scenes using Mask R-CNN and STIPS," in *International Conference on Machine Vision (ICMV)*, A. Verikas, D. P. Nikolaev, P. Radeva, and J. Zhou, Eds., vol. 11041, International Society for Optics and Photonics. SPIE, 2019, pp. 501–508.

[73] K. Host, M. Ivašić-Kos, and M. Pobar, "Tracking handball players with the deepsort algorithm," in *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 01 2020, pp. 593–599.

[74] Y. Guo, S. Lao, and L. Bai, "Player detection algorithm based on color segmentation and improved camshift algorithm," in *International Conference on Information Technology and Software Engineering*, W. Lu, G. Cai, W. Liu, and W. Xing, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 765–772.

[75] C. Brumann, M. Kukuk, and C. Reinsberger, "Evaluation of open-source and pre-trained deep convolutional neural networks suitable for player detection and motion analysis in squash," *Sensors*, vol. 21, no. 13, 2021.

[76] M. Radhakrishnan, "Human object detection and tracking using background subtraction for sports applications," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, p. 4, 05 2015.

[77] S. Lotfi, D. Aski, and S. Zakeri, "Soccer player detection and tracking based on image processing," *Acta Technica CSAV (Ceskoslovensk Akademie Ved)*, vol. 63, pp. 309–314, 01 2018.

[78] U. M. Rao and U. C. Pati, "A novel algorithm for detection of soccer ball and player," in *International Conference on Communications and Signal Processing (ICCSP)*. IEEE, Apr. 2015, pp. 344–348.

[79] Y. Yang and D. Li, "Robust player detection and tracking in broadcast soccer video based on enhanced particle filter," *Journal of Visual Communication and Image Representation*, vol. 46, 03 2017.

[80] R. Miyamoto and T. Oki, "Soccer player detection with only color features selected using informed haar-like features," in *Advanced Concepts for Intelligent Vision Systems - 17th International Conference, ACIVS 2016, Proceedings*, ser. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), C. Distante, D. Popescu,

P. Scheunders, W. Philips, and J. Blanc-Talon, Eds. Springer Verlag, Jan. 2016, pp. 238–249.

[81] U. R. M. and U. C. Pati, "A novel algorithm for detection of soccer ball and player," in *International Conference on Communications and Signal Processing (ICCSP)*, 2015, pp. 0344–0348.

[82] G. Zhu, C. Xu, Q. Huang, and W. Gao, "Automatic multi-player detection and tracking in broadcast sports video using support vector machine and particle filter," in *International Conference on Multimedia and Expo*. IEEE, 2006, pp. 1629–1632.

[83] S. Maćkowiak, M. Kurc, J. Konieczny, and P. Maćkowiak, "A complex system for football player detection in broadcasted video," in *International Conference on Signals and Electronic Circuits (ICSES)*, 2010, pp. 119–122.

[84] X. Bai, T. Zhang, C. Wang, A. Abd El-Latif, and X. Niu, "A fully automatic player detection method based on one-class svm," *IEICE Transactions on Information and Systems*, vol. E96.D, pp. 387–391, 02 2013.

[85] C. Cui, "Player detection based on support vector machine in football videos," *International Journal of Performability Engineering*, vol. 14, 02 2018.

[86] M. Koshkina, H. Pidaparthy, and J. H. Elder, "Contrastive learning for sports video: Unsupervised player classification," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2021, pp. 4528–4536.

[87] H. Pidaparthy, M. H. Dowling, and J. H. Elder, "Automatic play segmentation of hockey videos," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2021, pp. 4585–4593.

[88] M. Şah and C. Direkoğlu, "Review and evaluation of player detection methods in field sports," *Multimedia Tools and Applications*, Jun 2021. [Online]. Available: https://doi.org/10.1007/s11042-021-11071-z

[89] M. Buric, M. Ivasic-Kos, and M. Pobar, "Player tracking in sports videos," in *International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, Dec. 2019, pp. 334–340.

[90] N. Rahmad, N. A. J. Sufri, N. Muzamil, and M. A. As'ari, "Badminton player detection using faster region convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, pp. 1330–1335, 06 2019.

[91] J. Komorowski, G. Kurzejamski, and G. Sarwas, "Footandball: Integrated player and ball detector," *CoRR*, vol. abs/1912.05445, 2019. [Online]. Available: http://arxiv.org/abs/1912.05445

[92] S. Hurault, C. Ballester, and G. Haro, "Self-supervised small soccer player detection and tracking," in *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, ser. MMSports '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 9–18. [Online]. Available: https://doi.org/10.1145/3422844.3423054

[93] I. Biliškov, M. Šarić, M. Russo, and M. Stella, "Players detection using u-net based fully convolutional network," in *International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2021, pp. 1–5.

[94] M. Şah and C. Direkoğlu, "Evaluation of image representations for player detection in field sports using convolutional neural networks," in *International Conference on Theory and Application of Fuzzy Systems and Soft Computing (ICAFS)*. Cham: Springer International Publishing, 2019, pp. 107–115.

[95] X. You, D. Li, and M. Zhang, "A top-view multiple people tracking system based on newest yolov5 and deepsort using depth data," in *International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2022, pp. 91–96.

[96] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020.

[97] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *CoRR*, vol. abs/1703.07402, 2017. [Online]. Available: http://arxiv.org/abs/1703.07402

[98] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[99] M. Buric, M. Pobar, and M. Ivašić-Kos, "Adapting yolo network for ball and player detection," in *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 01 2019, pp. 845–851.

[100] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," *CoRR*, vol. abs/1602.00763, 2016.

[101] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[102] L. Zhang, Y. Lu, G. Song, and H. Zheng, "Rc-cnn: Reverse connected convolutional neural network for accurate player detection," in *Trends in Artificial Intelligence*, X. Geng and B.-H. Kang, Eds. Cham: Springer International Publishing, 2018, pp. 438–446.

[103] R. Theagarajan, F. Pala, X. Zhang, and B. Bhanu, "Soccer: Who has the ball? generating visual analytics and player statistics," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2018.

[104] K. Lu, J. Chen, J. J. Little, and H. He, "Light cascaded convolutional neural networks for accurate player detection," *CoRR*, vol. abs/1709.10230, 2017. [Online]. Available: http://arxiv.org/abs/1709.10230

[105] A. Senocak, T.-H. Oh, J. Kim, and I. So Kweon, "Part-based player identification using deep convolutional representation and multi-scale pooling," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, June 2018.

[106] J. Komorowski, G. Kurzejamski, and G. Sarwas, "Deepball: Deep neural-network ball detector," *CoRR*, vol. abs/1902.07304, 2019.

[107] Y. Zhao, C. Lei, and J. C. Patterson, "A piv measurement of the natural transition of a natural convection boundary layer," *Springer-Experiments in Fluids*, vol. 56, pp. 1891–5, 2015.

[108] V. L. Erickson, M. A. Carreira-Perpiñán, and A. E. Cerpa, "Occupancy modeling and prediction for building energy management," *ACM Trans. Sen. Netw.*, vol. 10, no. 3, may 2014. [Online]. Available: https://doi.org/10.1145/2594771

[109] J. R. Dobbs and B. M. Hencey, "Model predictive hvac control with online occupancy model," *Energy and Buildings*, vol. 82, pp. 675 – 684, 2014.

[110] B. Dong, B. Andrews, K. L. Poh, M. Höynck, R. Zhang, Y.-S. Chiou, and D. Benitez, "An information technology enabled sustainability test-bed (itest) for occupancy detection through an environmental sensing network," *Energy and Buildings*, vol. 42, no. 7, pp. 1038 – 1046, 2010.

[111] C. Jiang, M. K. Masood, Y. C. Soh, and H. Li, "Indoor occupancy estimation from carbon dioxide concentration," *Energy and Buildings*, vol. 131, pp. 132 – 141, 2016.

[112] H. Han, K.-J. Jang, C. Han, and J. Lee, "Occupancy estimation based on co2 concentration using dynamic neural network model," 09 2013.

[113] B. Dong and K. P. Lam, "Building energy and comfort management through occupant behaviour pattern detection based on a large-scale environmental sensor network," *Journal of Building Performance Simulation*, vol. 4, no. 4, pp. 359–369, 2011.

[114] T. H. Pedersen, K. U. Nielsen, and S. Petersen, "Method for room occupancy detection based on trajectory of indoor climate sensor data," *Building and Environment*, vol. 115, pp. 147–156, Jan. 2017.

[115] X. Liang, T. Hong, and G. S. Qiping, "Occupancy data analytics and prediction: A case study," *Building and Environment*, vol. 102, pp. 179 – 192, 2016.

[116] y. Benezeth, h. Laurent, B. Emile, and c. Rosenberger, "Towards a sensor for detecting human presence and characterizing activity," *Energy and Buildings*, vol. 43, no. 2, pp. 305 – 314, 2011.

[117] H.-C. Shih, "A robust occupancy detection and tracking algorithm for the automatic monitoring and commissioning of a building," *Energy and Buildings*, vol. 77, pp. 270 – 280, 2014.

[118] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Water filling: Unsupervised people counting via vertical kinect sensor," in *International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 215–220.

[119] R. Gade, A. Jørgensen, and T. B. Moeslund, "Long-term occupancy analysis using graph-based optimisation in thermal imagery," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 3698–3705.

[120] R. Gade, A. Jørgensen, and T. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," *International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, 01 2012.

[121] C. G. del Postigo, J. Torres, and J. M. Menéndez, "Vacant parking area estimation through background subtraction and transience map analysis," *IET Intelligent Transport Systems*, vol. 9, no. 9, pp. 835–841, 2015.

[122] R. S. B. Karunamoorthy and N. JayaSudha, "Design and implementation of an intelligent parking management system using image processing," vol. 4, no. 1, 2015.

[123] G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, and C. Vairo, "Deep learning for decentralized parking lot occupancy detection," *Expert Systems with Applications*, vol. 72, 10 2016.

[124] H. Chandel and S. Vatta, "Occlusion Detection and Handling: A Review," *International Journal of Computer Applications*, vol. 120, no. 10, pp. 33–38, 2015.

[125] T. Rathnayake, A. K. Gostar, R. Hoseinnezhad, R. Tennakoon, and A. Bab-Hadiashar, "On-line visual tracking with occlusion handling," *Sensors (Switzerland)*, vol. 20, no. 3, pp. 1–23, 2020.

[126] Q. Zhao, W. Tian, Q. Zhang, and J. Wei, "Robust Object Tracking Method Dealing with Occlusion," in *International Conference on Soft Computing and Machine Intelligence, ISCMI 2016*. Institute of Electrical and Electronics Engineers Inc., oct 2017, pp. 143–147.

[127] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving Occlusion and Hard Negative Handling for Single-Stage Pedestrian Detectors," *Conference on Computer Vision and Pattern Recognition*, no. i, pp. 966–974, 2018.

[128] H. Kataoka, S. Ohki, K. Iwata, and Y. Satoh, "Occlusion Handling Human Detection with Refocused Images," in *International Conference on Pattern Recognition*. IEEE, 2018, pp. 1701–1706.

[129] M. I. Shehzad, Y. A. Shah, Z. Mehmood, A. W. Malik, and S. Azmat, "K-means based multiple objects tracking with long-term occlusion handling," *IET Computer Vision*, vol. 11, no. 1, pp. 68–77, feb 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2016.0156

[130] J. Zhang, H. Sun, W. Guan, J. Wang, Y. Xie, and B. Shang, "Robust human tracking algorithm applied for occlusion handling," in *International Conference on Frontier of Computer Science and Technology*, 2010, pp. 546–551.

[131] J. Marín, D. Vázquez, A. M. López, J. Amores, and L. I. Kuncheva, "Occlusion handling via random subspace classifiers for human detection," *IEEE Transactions on Cybernetics*, vol. 44, no. 3, pp. 342–354, 2014.

[132] C. Zhou and J. Yuan, "Bi-box Regression for Pedestrian Detection and Occlusion Estimation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11205 LNCS, pp. 138–154, 2018.

[133] M. Babaee, Z. Li, and G. Rigoll, "Occlusion handling in tracking multiple people using rnn," in *International Conference on Image Processing, ICIP*. IEEE, 2018, pp. 2715–2719.

[134] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1815–1821.

[135] M. M. N. Ali, M. Abdullah-Al-Wadud, and S. L. Lee, "Multiple object tracking with partial occlusion handling using salient feature points," *Information Sciences*, vol. 278, pp. 448–465, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2014.03.064

[136] X. Lu, J. Zhang, L. Song, R. Lei, H. Lu, and N. Ling, "Person tracking with partial occlusion handling," in *Workshop on Signal Processing Systems (SiPS)*. IEEE, 2015, pp. 1–6.

[137] M. Thu and N. Suvonvorn, "Pyramidal Part-Based Model for Partial Occlusion Handling in Pedestrian Classification," *Advances in Multimedia*, vol. 2020, 2020.

[138] A. Sanin, C. Sanderson, and B. C. Lovell, "Improved shadow removal for robust person tracking in surveillance scenarios," in *International Conference on Pattern Recognition*, 2010, pp. 141–144.

[139] A. Russell and J. J. Zou, "Moving shadow detection based on spatial-temporal constancy," in *International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2013, pp. 1–6.

[140] T. F. Y. Vicente, M. Hoai, and D. Samaras, "Leave-one-out kernel optimization for shadow detection and removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 682–695, 2018.

[141] M. Kristo, M. Ivasic-Kos, and M. Pobar, "Thermal Object Detection in Difficult Weather Conditions Using YOLO," *IEEE Access*, vol. 8, pp. 125 459–125 476, 2020.

[142] P. Tumas, A. Nowosielski, and A. Serackis, "Pedestrian Detection in Severe Weather Conditions," *IEEE Access*, vol. 8, pp. 62 775–62 784, 2020.

[143] H. Kim, E. Chae, G. Jo, and J. Paik, "Fisheye lens-based surveillance camera for wide field-of-view monitoring," in *International Conference on Consumer Electronics (ICCE)*. IEEE, 2015, pp. 505–506.

[144] H. Kim, J. Jung, and J. Paik, "Fisheye lens camera based surveillance system for wide field of view monitoring," *Optik*, vol. 127, no. 14, pp. 5636–5646, 2016.

[145] T. Wang and C.-H. Liao, "People detection in downward-viewing fisheye camera networks using fuzzy integral," in *International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, June 2019, pp. 1–5.

[146] D. Levi and S. Silberstein, "Tracking and motion cues for rear-view pedestrian detection," in *International Conference on Intelligent Transportation Systems*. IEEE, Sep. 2015, pp. 664–671.

[147] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[148] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer, 2016, pp. 21–37.

[149] A. Cioppa, A. Deliège, and M. Van Droogenbroeck, "A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE/CVF, 2018, pp. 1846–184 609.

[150] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 580–587.

[151] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 1–9.

[152] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[153] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Workshops on Applications of Computer Vision*, vol. 1. IEEE, Jan. 2005, pp. 364–369.

[154] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 162–182, 2007.

[155] R. Miezianko and D. Pokrajac, "People detection in low resolution infrared videos," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2008, pp. 1–6.

[156] Y. Socarras, S. Ramos, D. Vázquez, A. López, and T. Gevers, "Adapting pedestrian detection from synthetic to far infrared images," in *International Conference on Computer Vision (ICCV) Workshop*, 1 2013.

[157] C. Dai, Y. Zheng, and X. Li, "Layered representation for pedestrian detection and tracking in infrared imagery," in *Conference on Computer Vision and Pattern Recognition(CVPR) Workshops*. IEEE, Sep. 2005.

[158] W. Li, D. Zheng, T. Zhao, and M. Yang, "An effective approach to pedestrian detection in thermal imagery," in *International Conference on Natural Computation*, May 2012, pp. 325–329.

[159] P. Tumas, A. Jonkus, and A. Serackis, "Acceleration of hog based pedestrian detection in fir camera video stream," in *Conference of Electrical, Electronic and Information Sciences (eStream)*. IEEE, 2018, pp. 1–4.

[160] D. Heo, E. Lee, and B. C. Ko, "Pedestrian detection at night using deep neural networks and saliency maps," *Electronic Imaging*, vol. 2018, no. 17, pp. 060 403–1, 2018.

[161] C. Herrmann, T. Müller, D. Willersinn, and J. Beyerer, "Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs," in *Electro-Optical and Infrared Systems: Technology and Applications XIII*, D. A. Huckridge, R. Ebert, and S. T. Lee, Eds., vol. 9987, International Society for Optics and Photonics. SPIE, 2016, p. 99870I. [Online]. Available: https://doi.org/10.1117/12.2240940

# Chapter 2

# Summary of research and contributions

As mentioned in the previous chapter, this thesis consists of four main technical parts (II-V) documenting the research conducted to achieve the project goal of occupancy analysis using diverse camera setups. The summary of each of those studies, contributions and findings are provided in this chapter.

## 1 Chapter 3: Experimental setup to investigate the feasibility of long-term data recordings in outdoor environment

This chapter reports the "Experimental setup to investigate the feasibility of long-term data recordings in an outdoor environment". This study aims to define and evaluate different camera setups based on the feasibility of installation, cost, area coverage and other factors. Three camera types, i.e. fish eye, wide angle action and thermal, are used in this study. With different combinations, five camera setups are initially proposed, keeping in mind the ease of installation. The suggested camera setups are considered mainly due to their coverage area. Two camera setups, namely, 1) one fisheye and 2) two wide-angle cameras covering the whole field, are the simplest solution. Another setup consisting of three thermal cameras is also proposed due to its better performance in low light and total dark conditions. To reduce the complexity of multiple camera installations, another setup consisting of only one thermal panning camera is investigated. The last proposed setup combines one fisheye and one thermal camera. Although the installation setup

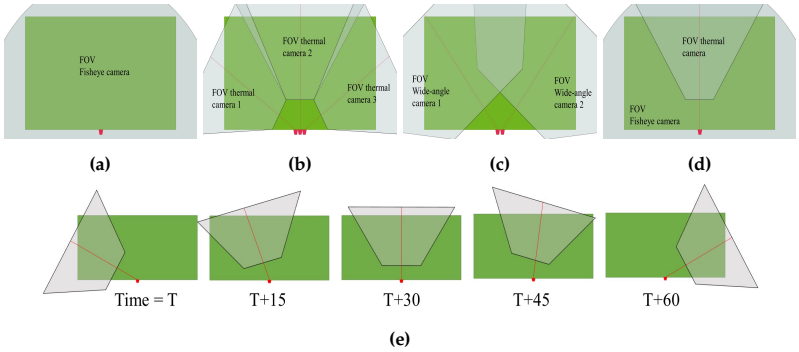



**Fig. 2.1:** A grey area shows the coverage area while the green area represents the field. The red dot indicates the position of the camera. **(a)** One fisheye camera, **(b)** Three thermal cameras, **(c)** Two wide-angle cameras, **(d)** One fish eye and one thermal camera, **(e)** Rotating thermal camera at different time instances.

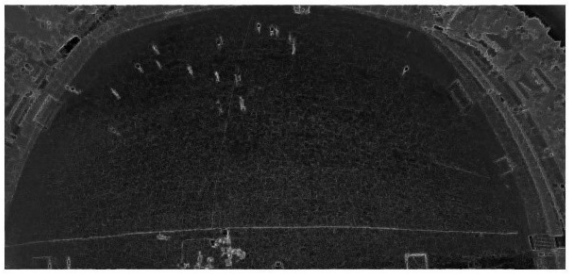
seems complex, the aim is to achieve day and night full field coverage. The fields of view of the five setups are shown in figure 2.1.

One day data for each setup was recorded at a soccer field in Aalborg, Denmark, on the same day. Fortunately, the recorded data was diverse, having all the variations of sunlight, the shadow at different angles, rain, wind, night time and shadows cast by moving clouds. The video recordings were then thoroughly analyzed in terms of camera cost, recognition visibility, day/night vision, occlusion, complexity, weather effect, and area coverage. Finally, the study deduced that the setups with two wide-angle cameras and one rotating thermal camera could not be tested further due to their limitation, including power, storage and full-time area coverage, respectively.

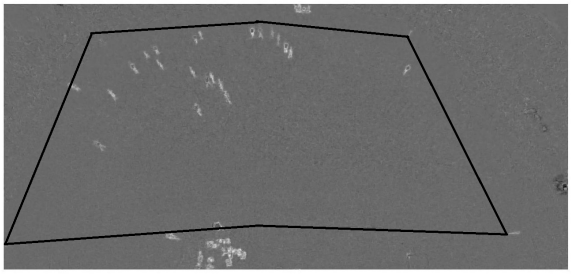
## 2 Chapter 4: Occupancy Analysis of Soccer Fields Using Wide-Angle Lens

This chapter investigates the "Occupancy Analysis of Soccer Fields Using Wide-Angle Lens" [1]. A camera setup with one fisheye is used in this study, and the aim is to perform player detection-based occupancy analysis over a specific interval of time. Data analysis revealed a few challenges, i.e. a) image distortion around the corners and b) players appear very small, especially as they move away from the camera. Furthermore, the analysis showed that detecting occluded and blur players is hard. To address these issues, this paper first deals with distortion correction by implementing barrel distortion correction presented in [2]. Afterwards, image enhancement is performed by implementing 2-D wavelets to deal with blurring and enhancement of field

area, explicitly enhancing the directional edges in the image.



**(a)**



**(b)**

**Fig. 2.2: (a)** Image after enhancement, **(b)** Image after background subtraction [1].

The enhancement of the images leads to the sharp appearance of players. However, noise from distortion or other edgy objects in the field, like the lines or the grass, is also enhanced. In order to suppress it, background subtraction is employed, which significantly improves the image quality (see, Fig 2.2. Other small noise pixels are removed by using morphological operations. This leads to our candidate player regions in the image. Each candidate region is further categorized using the threshold method based on compactness and colour information to cater for occlusion. Six thousand frames are manually labelled as ground truth to calculate the results. The work has achieved a small average error of 2.67% in the full activity period, 3.64% during the transition period and o.00096 % in no activity period. Fig 2.3 shows the results.

# 3 Chapter 5: Estimating the Number of Soccer Players using Simulation-based Occlusion Handling

This chapter, "Estimating the Number of Soccer Players using Simulation-based Occlusion Handling" [3] also focuses on occupancy analysis but using

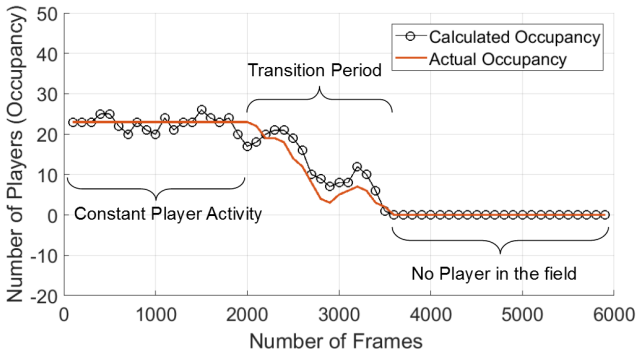



**Fig. 2.3:** Occupancy analysis over 6000 samples [1].

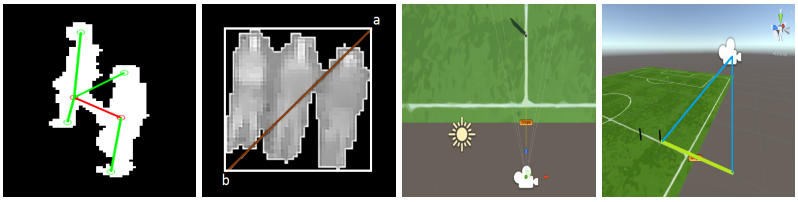



**Fig. 2.4: (a)** red circles show the connected points whereas green circles and lines are the branch points to calculate connected point distance and slope, **(b)** The diagonal length of a bounding box is the length of red line between the corners **a** and **b**, **(c)** Top view of the virtual setup, **(d)** Side view of the virtual setup [3].

a thermal camera. Thermal cameras provide better data quality, especially in low and no-light conditions. However, it is challenging to cater for occlusion, especially for objects away from the camera, because of less textural and no colour information. Therefore, the paper introduces a method that combines machine learning and virtual reality-based estimation of the number of players in the field to perform occupancy analysis. In the proposed method, candidate player regions are first identified using maximum entropy-based thresholding. The thresholding segments the light player regions from the dark background. Further, morphological operations are utilized to remove the small noise blobs. Occluded and non-occluded players are then classified using bragged tree classifier [4], where connected point-slope, connected point distance (see, Fig 2.4a), convex area and finally the diagonal length of bounding boxes (see, Fig. 2.4b) are used as input features.

Simulations on Unity 3D [5] are performed by making a virtual setup for identifying the number of players, i.e., 2, 3 and 4, in the occluded blob (see Fig. 2.4c, Fig. 2.4d). The virtual setup recreates the occlusions by making

one player still and moving the other players from right to left in a loop. The program measures 9978 possible instances for two-player occlusions, 12401 possibilities for three occluded players and 33001 possibilities for four. Afterwards, the number of players is identified using maximum likelihood-based density estimation, in which blob sizes in simulations are compared with the blob sizes of the original image.

The tests are performed on 8990 frames from a video of five minutes containing 71443 players. The paper has achieved an accuracy of 96.1% in detection and a precision of 97.8% in carrying out the occupancy analysis (see Fig.2.5).

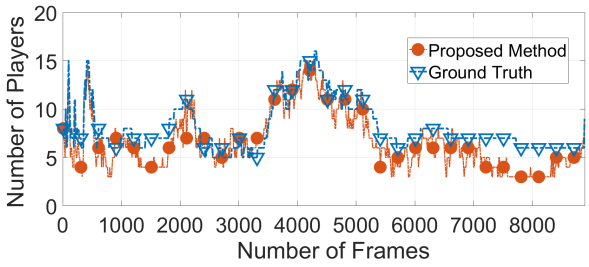
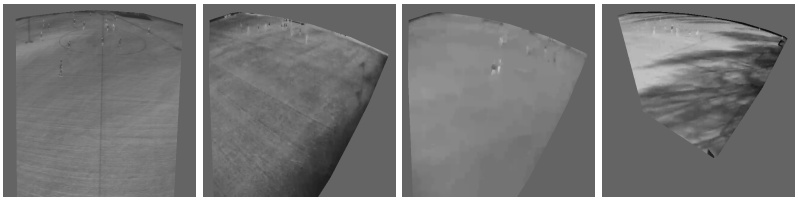



**Fig. 2.5:** Orange line represents the estimated number of people and the blue line shows the ground truth [3].

# 4 Chapter 6: The Effect of a Diverse Dataset for Transfer Learning in Thermal Person Detection

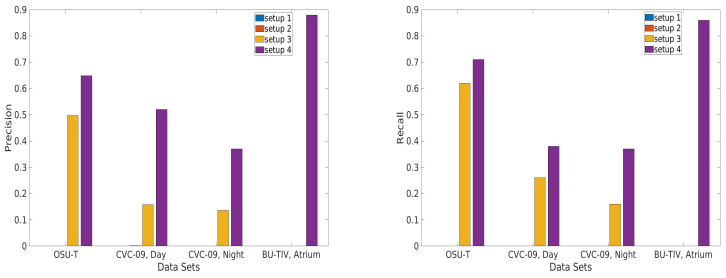
To further study the thermal performance using deep learning, long-time data analysis is performed in "The Effect of a Diverse Dataset for Transfer Learning in Thermal Person Detection" [6]. The paper thoroughly reviews available thermal datasets for person detection in thermal. Unfortunately, the online available thermal datasets have a person visibility of more than 50% and, in most cases, more than 70%. This makes it impossible to use these datasets and adapt the model for real-life application areas. The twenty-week thermal dataset is recorded and analyzed to address this challenge. The study indicated that the weather, light and environmental effects are more visible in thermal than in RGB camera setup (Fig. 2.6).

Transfer learning saves the cost of annotating a tremendous amount of training data for deep neural networks. The paper uses YOLO v3 network [7] as a base model, and two-step transfer learning is performed. In the first step, human features are learned from the RGB dataset. In the second step, online



(a) (b) (c) (d)

**Fig. 2.6:** Some challenging characteristics in thermal data. (**a**) Varying body temperatures. (**b**) Similar temperatures. (**c**) Motion blur due to wind. (**d**) Shadows [6].




(a) (b)

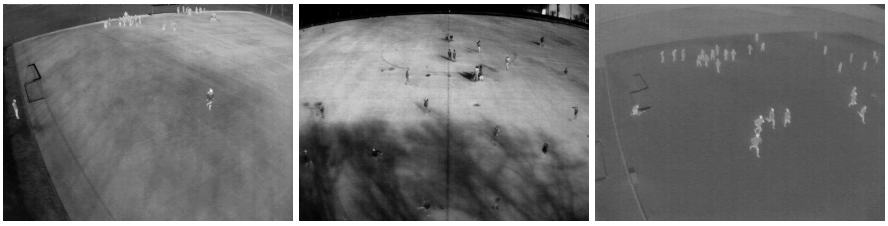
**Fig. 2.7:** (**a**) Precision and (**b**) recall measures of different training weights on publicly available datasets. Here, the blue bars are the results tested by our thermal training weights, and orange bars are the results tested by our thermal training weights and further training by adding only 5% of the new dataset for 100 iterations [6].

available thermal data with good person visibility is used to make the network learn human detection features. Later the diverse data with all possible outdoor scenarios are added to the training data to let the network adapt to our real-life application. The diverse outdoor data is categorized into nine phenomena, i.e., good condition, far viewpoint, opposite temperature, similar temperature, low resolution, occlusion, shadow, snow, and wind. The effect of each phenomenon is investigated by adding data from different phenomena and then combinations of phenomena and observing the results on testing data. The testing data consists of 1000 randomly selected images from twenty weeks of data recordings. Furthermore, the trained network with the best combination of data is also tested with publicly available datasets (Fig. 2.7).

The paper suggests that each data category affects the performance differently for a given application. The data categories could be selected intelligently per the application requirements to save the time and effort of data annotation. The diverse data set is publicly available for further research, and investigation [8].

# 5 Chapter 7: Effects of Pre-processing on the Performance of Transfer Learning Based Person Detection in Thermal Images

Further studies in the thermal domain to investigate the role of different polarities and pre-processing techniques are investigated in "Effects of Pre-processing on the Performance of Transfer Learning Based Person Detection in Thermal Images" [9]. This study is the continuation of [6] to analyze the behaviour of different phenomena data. Thermal data typically possess two main polarities, i.e., light-person representation on dark background and dark-person representation on a light background. In the previous study, [6], it was observed that results changed significantly whenever the data from opposite polarity was added. In this paper, we focused on improving results for such images, i.e. similar temperature (minimum contrast between person representation and background) images and diverse polarity images (Fig. 2.8). To cater to similar temperature images, we implemented contrast enhancement. In addition, an automated process is proposed for diverse polarity images to detect opposite polarity images and convert them to the same polarity before further processing.



**(a)** **(b)** **(c)**

**Fig. 2.8:** **(a)** Person appeared similar to background, **(b)** Person appeared darker w.r.t background and **(c)** Person appeared lighter w.r.t background [9].

Contrast enhancement is performed by using histogram equalization. Whereas opposite polarity detection is performed in two steps, i.e., 1) sunlight detection and 2) human body temperature detection. High sunlight is detected by summing up the lighter pixels in an image, while the human body temperature is detected by studying the histogram bins of the high-intensity images. The proposed method of studying the effect of data homogenization in single polarity is evaluated using YOLO v3 [7]. It was observed that the best results are obtained without applying pre-processing techniques or homogenization and by using the data in its original form.

(a) (b) (c)

**Fig. 2.9: (a)** Projection of the thermal image onto the fisheye image. The thermal camera sees only $\approx$ 22% of the football field pixels of the fisheye image. **(b)** The bounding boxes given by the teacher network. **(c)** surrogate ground-truth bounding boxes [10].

# 6 Chapter 8: Multimodal and multiview distillation for real-time player detection on a football field

Player detection and occupancy analysis using one fish eye and one thermal camera is presented in "Multimodal and multiview distillation for real-time player detection on a football field" [10]. In this setup thermal camera only covers the middle of the field, while the fisheye camera covers the whole field. The main objective of this setup is to achieve full area coverage during the day and night. Moreover, the setup is simple and cost-effective in installation. In this work, the idea is to cross-learn from thermal to fisheye and then transfer the learning parameters to non-thermal coverage parts of the field. Data preparation is an essential step in such a setup. As both cameras have different clock cycles and camera parameters, time synchronization and model transfer parameters for spatial transformation or mapping of pixels between two cameras are performed as a foremost step before further processing (Fig. 2.9a).

The online distillation method, based on a teacher-student network [11], is applied to learn from thermal to fisheye. As a teacher network, YOLO is used, previously trained on thermal data [10], to detect the players in thermal (Fig. 2.9b). Afterwards, the bounding boxes are transferred from thermal to fisheye image using camera calibration and model transfer parameters. This gives some surrogate ground truth annotation (Fig. 2.9c) for retraining the fisheye image for the thermal-fisheye overlapping area using tinyYOLO. To extend the training data to a non-overlapping thermal-fisheye region in the fisheye image, we have augmented the players by copying and pasting the player from the overlapping thermal-fisheye region to the non-overlapping thermal-fisheye region. As distortion affects differently at different parts of the fisheye image, depending on the distance from the camera, the player regions are first

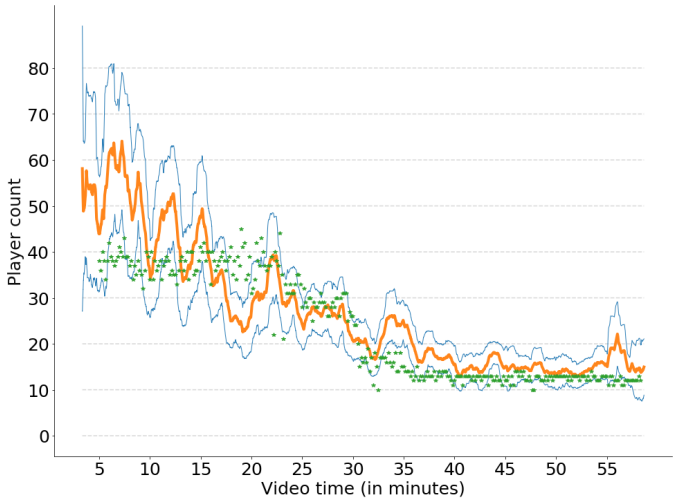



**Fig. 2.10:** **Results on the player counting task** averaged over a 1-minute window, and associated **standard deviation** [10].

scaled, rotated, and blended before pasting to match the specifications of the particular region in the fisheye image. In this manner, some fake ground truths are generated. To solve the problem of initially present players in non-overlapping regions, we used viBe [12] background subtraction to remove the original players while training and only keep the augmented players. The results show that the teacher-student network improves over time. It is noticed that the network achieved RMSE of 3.4 players (Fig. 2.10) w.r.t ground truth. The code and data are also made publicly available for further research [13].

# 7 Chapter 9: Report to municipality

The chapter covers the aspects of the final report to the Aalborg municipality. It goes through the specifications of the final setup used for occupancy analysis of the outdoor soccer fields. It also covers the algorithm details, evaluation methods, limitations in the detection and the format of the final report that is presented to the municipality.

# 8 Contributions

This section will sum up the contributions made in this thesis.

- **Analysis of different test camera setups** for occupancy analysis in an

outdoor environment is performed in chapter 3. The analysis investigated the requirements for the occupancy analysis application in Denmark's local soccer field. A comparison of different test setups is also presented in the chapter.

- Occupancy analysis of outdoor soccer field using a single **fisheye camera** is presented in chapter 4. Method for occupancy analysis based on player detection is presented, where player detection is performed using a features-based machine learning algorithm.
- Occupancy analysis of outdoor soccer field using **thermal camera** is presented in chapter 5. In this setup, player detection is based on a features-based machine learning algorithm, and occlusion is handled using simulation-based computer graphics.
- Diverse **thermal dataset** is presented to research community. Chapter 6 explains the dataset and the findings about the need for diversity in a dataset. This chapter also explains how different phenomena in thermal data affect the learning of a CNN.
- Occupancy analysis using dual modalities, i.e. **one fisheye and one thermal camera** is presented in chapter 9. The research deals with the multiview distillation problem. Point-to-point registration is performed for a common view, and then knowledge is transferred using student-teacher-based network training approach [11].
- **Fisheye-thermal Dataset** for cross-model learning between fisheye and the thermal camera is also presented in chapter 9 for further studies and research.
- chapter 7 rules out the need to perform data homogenization in **thermal images** before feeding to Deep Neural Network. A machine-learning method for high sunlight detection in thermal data is also presented.

# References

[1] N. U. Huda, B. D. Hansen, R. Gade, and T. B. Moeslund, "Occupancy analysis of soccer fields using wide-angle lens," in *International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, 2017, pp. 354–359.

[2] D. Laxman and S. Bhandari, "Implementation of barrel distortion correction algorithm for wide angle camera based systems," *i-manager's Journal on Image Processing*, vol. 4, p. 43, 01 2017.

[3] N. U. Huda, K. H. Jensen, R. Gade, and T. B. Moeslund, "Estimating the number of soccer players using simulation-based occlusion handling," in *Conference on*

*Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2018, pp. 1905–190 509.

[4] L. Breiman, "Bagging predictors," *Machine Learning, Springer*, vol. 24, no. 2, pp. 123–140, 1996.

[5] Unity Technologies, "Unity," https://unity3d.com/unity.

[6] N. U. Huda, B. D. Hansen, R. Gade, and T. B. Moeslund, "The effect of a diverse dataset for transfer learning in thermal person detection," *Sensors*, vol. 20, no. 7, p. 1, Apr. 2020.

[7] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[8] N. U. Huda, B. D. Hansen, R. Gade, and T. B. Moeslund, "Person detection dataset (pd-t)," https://vap.aau.dk/dataset/.

[9] N. U. Huda, R. Gade, and T. B. Moeslund, "Effects of pre-processing on the performance of transfer learning based person detection in thermal images," in *International Conference on Pattern Recognition and Machine Learning (PRML)*. Chengdu, China: IEEE, 2021, pp. 86–91.

[10] A. Cioppa, A. Deliège, N. U. Huda, R. Gade, M. V. Droogenbroeck, and T. B. Moeslund, "Multimodal and multiview distillation for real-time player detection on a football field," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, June 2020.

[11] A. Cioppa, A. Deliège, M. Istasse, C. De Vleeschouwer, and M. V. Droogenbroeck, "ARTHuS: Adaptive Real-Time Human Segmentation in Sports Through Online Distillation," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, June 2019.

[12] O. Barnich and M. V. Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, June 2011.

[13] A. Cioppa, A. Deliège, N. U. Huda, R. Gade, M. V. Droogenbroeck, and T. B. Moeslund, "Multimodal and multiview distillation for real-time player detection on a football field," https://github.com/cioppaanthony/multimodal-multiview-distillation.

# Part II

# Preliminary Studies

# Chapter 3

# Experimental setup to investigate the feasibility of long-term data recordings in outdoor environment

## 1 Introduction

Data recording is an essential block in the pipeline of computer vision as it is the first and foremost step for every machine vision application [1]. The outcome, as well as the result of each step in the pipeline, depends on the quality of the data recorded. Our primary concern is to have image data covering the entire football field. Since the soccer is rather big, which sensor setup to select is not apparent. Moreover, our focus is long-term analysis, so we must deal with changing outdoor conditions. Also, the cost of the sensor suite is relevant to consider. This chapter defines and compares five different camera setups by using three cameras analytically. The criteria include complexity, price and image visibility in diverse weather and light conditions. The following section defines the required considerations.

## 2 Requirement considerations

Before initiating the study, some prior considerations are required to narrow the number of camera setups. Following are the considerations we took before employing a setup. It comprehends the essential demands of the setup to carry out the recording of twenty weeks.

## 2.1 Coverage

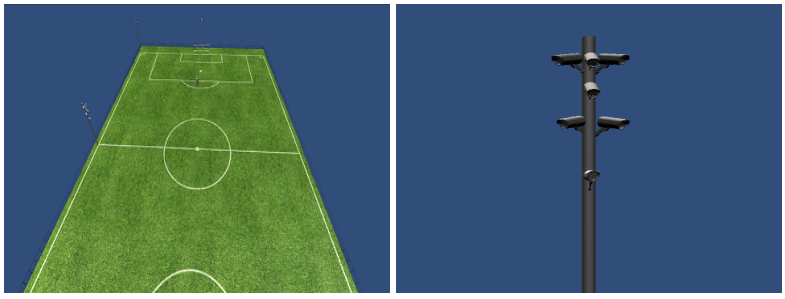
The project aims to cover the whole field area and to count the persons inside that area. The maximum field size that can be encountered is 105 x 68 m. Therefore, the camera setup should be able to cover the whole area and recognize the presence of any person within the specified area.

## 2.2 System simplicity

We intend to imply the camera setup for local football fields with fewer resources. It is hardly possible to set up at different locations with network availability. Manual cabling for the network at different locations could lead to disturbance in the play. We aim to deploy the setup to keep it simple in installation and integration with the existing setup. This ultimately means the sensor suite is mounted in only one location.

# 3 Proposed Camera Setups

Initial brainstorming rooted in the requirements and prior work resulted in five camera setups with three different types of cameras initially proposed to study further. All the setups are employed on a single pole to keep in mind the simplicity of installation. The height of the highest camera is kept at 9.8m from the ground. The other setups lie 50 cm below each other. Fig. 3.7 shows the setups for the day. Following are the proposed camera setups.



**(a)** **(b)**

**Fig. 3.1: (a)** Virtual view of proposed setup. **(b)**Closeup cameras. The three cameras at the top of the pole are thermal cameras. Just below is one fisheye camera. Below the fisheye camera are two wide-angle action cameras. The last one is the thermal panning camera.

## 3.1 Cameras

The lens concept was first explored 1000 years back in the 'Book of optics' (Kitab al manazir) by Ibn al-Haytham [2]. Later the idea of a camera setup was first introduced in 1021, and the cameras became commercialized in 1888 [3].

At the beginning of the camera revolution, cameras were not able to record images. Instead, a pinhole setup was used to project the light on another surface upside-down in greyscale. With the evolution of technology, it became possible to record the image on a film in analogue cameras and instant development inside the camera in Single Lens Reflex (SLR)-based technology. With more advancements, films get replaced by an imaging sensor that breaks the entering light into pixels and stores it in the form of numbers for each pixel which is the mechanism of Digital Single Lens Reflex (DSLR) cameras [3].

The possibilities of three main cameras are explored in this work. The considered camera setups are a combination of the mentioned cameras. The following sections provide a brief history and introduction of used cameras.

### Wide angle action camera

Many lenses and combinations have evolved over the centuries for photographing and recording [4]. For example, the wide-angle camera is made to maintain a small focal length for attaining a wide angle and increasing the size of the image. A small negative element is placed in front of the anterior focus of the lens projection to attain a large field of view. The setting with the combination of the lens and the large negative lens is called the reversed telephoto [4]. The setting makes it easier to capture magnified images by increasing the back focal distance of the lens.

We used wide-angle cameras specially designed for recording action videos. The cameras can capture vast horizontal and vertical angles and be made to record action sports and event videos [5].

### Fisheye Camera

Fisheye cameras are wide-angle RGB cameras, first introduced in 1906 by Wood [6]. The name fisheye is derived from the fact that the initial theory and development of the camera are based on the assumption of how we appear to the fish underwater. The lens development started in the 1920s.

It is impossible to capture an expansive view with a standard rectangular lens. Moreover, making a rectangular lens that can give coverage of more than 100° is challenging [7].

The lens of a fisheye camera is a convex-shaped, ultra-wide lens. It creates a spherical dome image comprising an extreme distortion due to its non-

linear distribution [8]. Two main kinds of lenses are common in the fisheye camera. One of them is a circular lens. The whole image dome is inscribed inside the film in the circular lens. The image is created by making horizontal, vertical and diagonal FOV of 100°. The other one is the full-frame fisheye lens in which the circular dome of the image lies outside the film, and it is obtained by diagonal coverage of 100°.

The objects captured by a fisheye lens vary in appearance depending on the distance from the camera. The objects just below the camera appear very large and convex. As the objects move away from the camera, they appear very small and distorted. Image correction may lead to significant loss of information and overall image quality [8].

Due to the viewing angle of more than 180°, the camera was initially utilized to monitor clouds and forests. Nowadays main application areas for fisheye lie in video surveillance. Other than surveillance, the cameras are also utilized robotics, satellite position other computer vision areas [9, 10].

Despite a severe distortion around the corners in the circular lens, a fish-eye camera provides extensive area coverage and achieves simplicity within the mid-price range.

**Thermal Cameras**

Thermal sensing devices were introduced mainly for military use in the 1940s and 1950s. The thermal sensing devices use the principles of infrared (IR) radiation to create images [11]. The devices become publically available in the 1980s.

Every object has some body temperature. Thermal sensing devices detect a heated object in terms of the temperature difference. The more hot an object is, the brighter it appears in a thermal image. However, the object can only be distinguished if its body temperature varies from the background and other objects [12]. Two primary imaging methods are used to capture the image. In the first method, the sensor detects the image either pixel by pixel or row by row. In the second method, the sensor detects the whole image simultaneously. The first method is more time-consuming and less popular nowadays [13].

Every camera detects electromagnetic radiation. As opposed to the other RGB cameras that detect the light reflected by the objects, thermal cameras absorb and detect the radiations emitted by an object. For thermal cameras, the heat radiations to be detected lie in the infrared region of the electromagnetic spectrum. The objects captured by thermal cameras lie in the range of black, white and different scales(256) of grey [13]. Primarily white represents the hot, but black as a heat source could be more understandable in some setups. Pseudo-colours may also be used to understand different heated objects in an image clearly.

The lens of a thermal camera is not made of glass, as ordinary glass blocks thermal radiation. Germanium and Chalcogenide glass(germanium-based) are mainly used as a lens. The materials let pass the thermal radiations and block the light [13].

Thermal cameras are also mainly used in surveillance, especially in night vision and low-light conditions. However, their applications are not limited to surveillance. They are widely used in industries, medicine, traffic and many more. [13].

## 3.2 Setups

The following setups are being studied by using the cameras mentioned above.

**One fisheye camera (FC)**

One Hikvision wide-angle Network Camera is used to capture the whole field in the test setup. The camera's field of view is kept 360° with the resolution of 1280 × 1280. The camera recorded the video at 10fps for ten hours. Data is transmitted by storing it on an SD card and then transferring it to the computer for further processing. Images are compressed in JPEG, and video is compressed in H264 format. The camera is derived by power over Ethernet. The camera setup and the image example are shown in Fig. 3.2.

**Three thermal cameras (TC)**

Three AXIS Q1922-E, Thermal Network Cameras inline cover the whole field. Each camera's horizontal field of view is 57° with a resolution of 640x480. The cameras record the videos at variable rates(29-30 fps) for 10 hours. Data is transmitted by recording on SSD and then transferring to a computer for further processing. For further study, the videos require time synchronization. Overlapping regions between different camera setups are segmented out. The images are stored in JPEG, and videos are stored in H264 format for further studies. The cameras derive on power over Ethernet. The setup does not cover the field area right below the camera poles. Fig. 3.3 shows the camera setup, coverage area, and example images.

**Two wide-angle action cameras (WC)**

Two Go pro 5 cameras are set up to record the field at 30fps for 10 hours. The camera will be mounted on a pole just below the fisheye camera. The cameras operate on a battery, and the recordings are captured in SSD storage. A small shield will be created to protect the cameras from rainwater. The video is

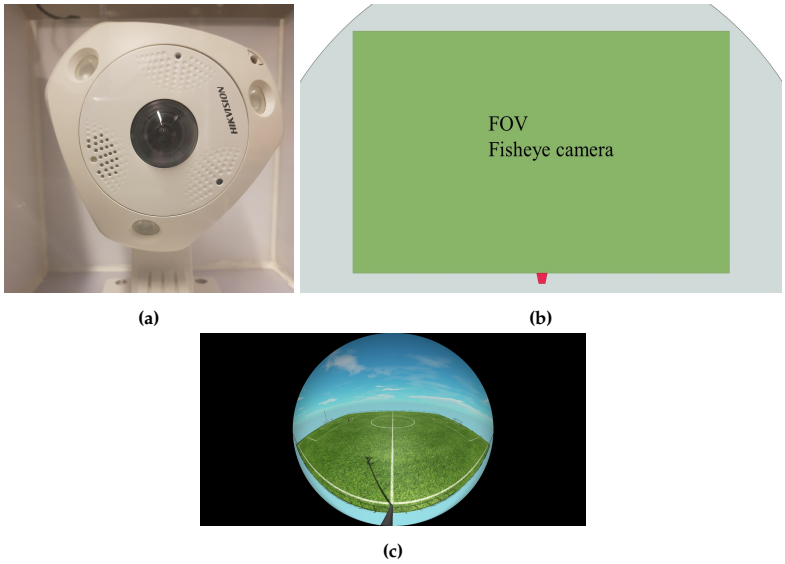



**Fig. 3.2: (a)** Hikvision Camera, **(b)** Field of view and area coverage of the camera. Where the green rectangle represents the field and a grey area represents the camera coverage, **(c)** Virtual image representation of field through the camera.

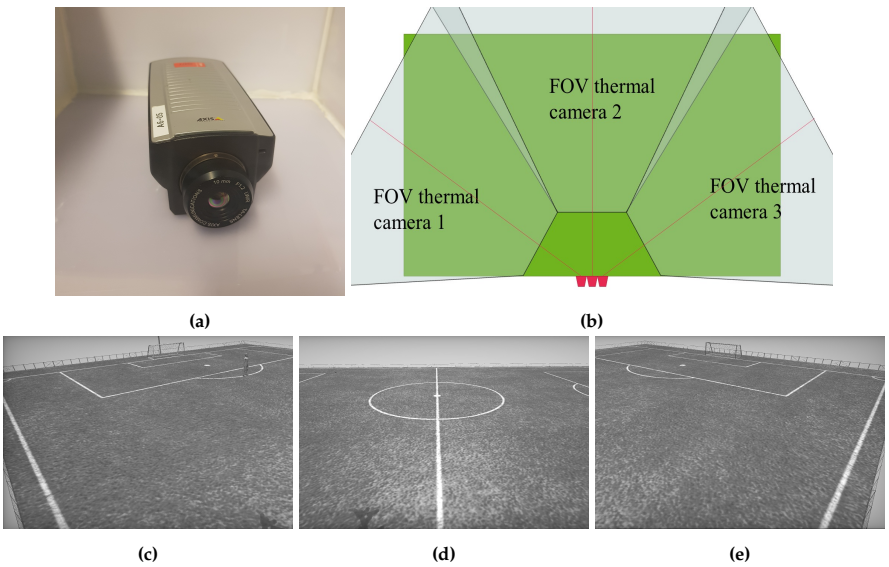



**Fig. 3.3: (a)** Thermal camera, **(b)** Field of view and area coverage of the camera. Where the green rectangle represents the field, and a grey area represents the camera coverage, **[(c),(d),(e)]** Virtual image representation of field through cameras 1, 2, and 3, respectively.

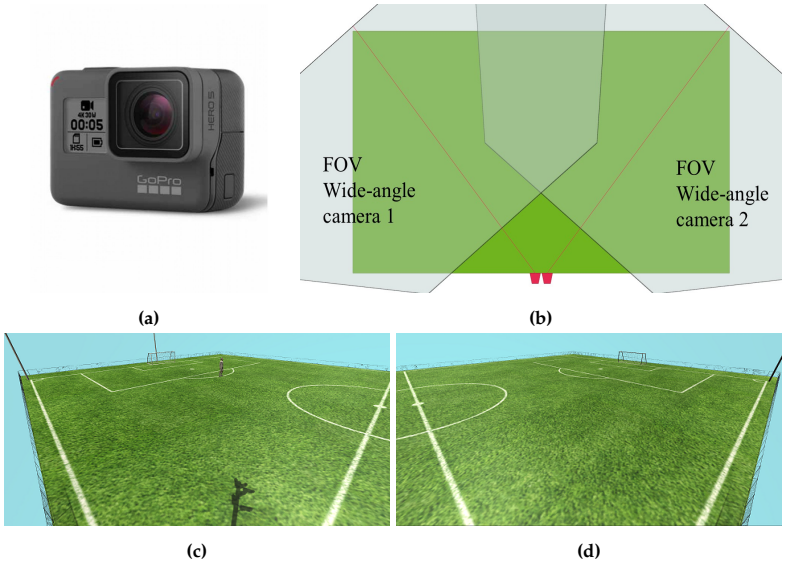



**Fig. 3.4: (a)** Wide angle action camera, **(b)** Field of view and area coverage of the camera. Where the green rectangle represents the field and a grey area represents the camera coverage, **[(c),(d)]** Virtual image representation of field through camera1 and camera2, respectively.

recorded as compressed in H264 formal. As the cameras can capture wide-angle horizontally and vertically, a small field area below the pole is left unattended. Fig. 3.4 shows the camera setup, coverage area, and example images.

### Multi-modal, One fisheye and one thermal camera (F-T-C)

The fisheye, in combination with the middle thermal camera, is studied to monitor the pros and cons of each camera. The recordings for both cameras are stored in an SSD. The time, space and frame rate synchronization is performed after recording. The setup gives a full view of the field. The camera setup, coverage area and example images are shown in Fig. 3.5.

### One Rotating/panning Thermal camera (RTC)

One thermal panning camera is used to study the outcome and possibility of monitoring the activities on the soccer field. The recordings are collected via storage on SSD for further study. The camera operated at 29fps with a resolution of 648x480. The camera setup and example frames are shown in Fig. 3.6.

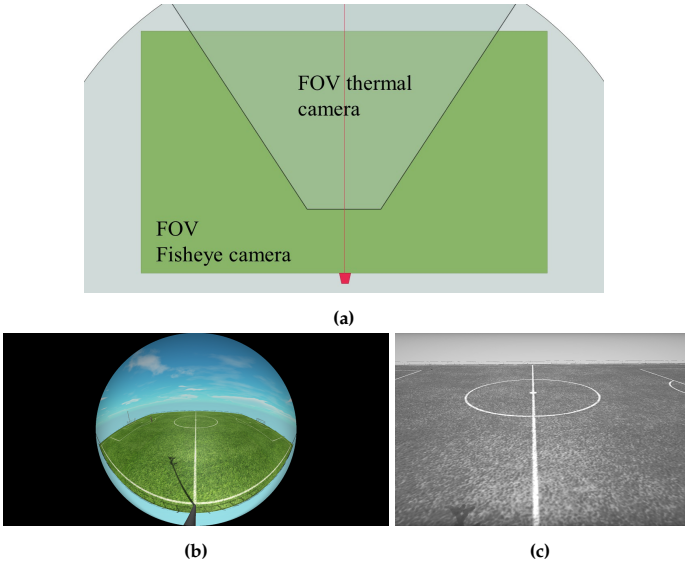



**Fig. 3.5: (a)** Field of view and area coverage of the F-T-C camera setup. Where the green rectangle represents the field, and a grey area represents the camera coverage, **(b)** Virtual image representation of field through the thermal camera, **(c)** Virtual image representation of field through the fisheye camera.

# 4 Pilot study-Behavioral Analysis

The setups were installed in a field in Aalborg, Denmark, on the 5th of October, 2017. All the cameras are installed for one day. Fig. 3.7 shows the setup. The outcome of the setups is thoroughly studied and observed in terms of resolution, complexity, price, occlusion, contrast, and image appearance in different light and weather conditions. Fortunately, the day depicted the overall weather and light conditions that a camera can encounter on a typical day in northern Denmark. That conditions mainly include wind, sun, shadows, and rain. In addition, cameras also captured daytime, nighttime, and the transition of day-to-night recordings. Following are the five proposed camera setups.

## 4.1 Initial observations

Following are the initial observations based on analyzing and studying the recorded data.

- The fisheye camera is effortless and cheap to employ, but the image quality could be better, especially at the corners of the fields. The identification of players becomes impossible in the corners, especially at

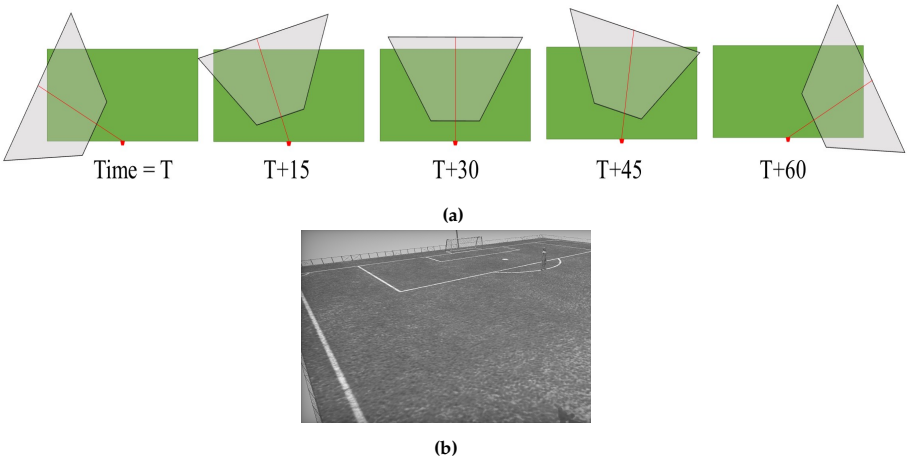



**Fig. 3.6: (a)** Field of view and area coverage of the RTC camera setup. Where the green rectangle represents the field and a grey area represents the camera coverage, **(b)** Virtual image representation of field through the camera.

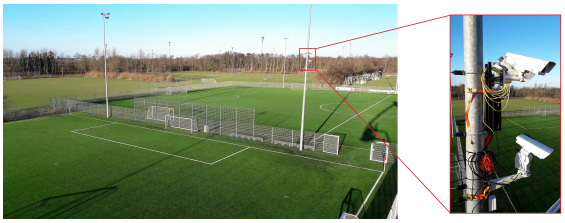


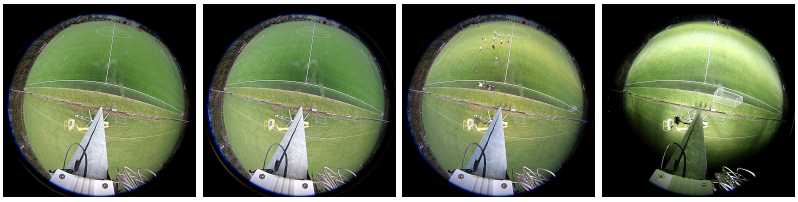
**Fig. 3.7:** Field to be captured along with the camera setups on the pole.

night, e.g., Fig. 3.8

- Wide-angle give a good contrast of players over green grass with a very good resolution. However, they also capture the clouds and different light intensities more clearly. Shades and intensity variations are more difficult to cater to in wide-angle. The effect of wind (pole shaking and displaced images) is also more prominent in wide-angle go pro cameras because of their small size. Fig. 3.9.

- Thermal cameras give a perfect result in the near field region. It is a good option to cater to light variation and weather effects. However, in far-field regions, player appearance is microscopic. This makes it difficult to cater for occlusion. Fig. 3.10.

- Player recognition and tracking is more difficult in thermal and fisheye cameras due to less resolution.



(a) (b) (c) (d)

**Fig. 3.8:** Example images captured by the fisheye camera. **(a)** and **(b)** are from daytime while **(c)** is from evening time and **(d)** is from night time.

(a) (b) (c) (d)

**Fig. 3.9:** Example images captured by wide-angle camera 1. **(a)** and **(b)** are from daytime while **(c)** is from evening time and **(d)** is from night time.

- Rain effect is not very visible in any of the cameras.

# 5 Comparison analysis

The parameters chosen for comparison can vary depending on the desired outcome. Real-world problem solving takes account of possible scenarios and readily available resources. Resource management and accessibility play an important role in finding solutions. Depending upon the municipality's requirement, local weather conditions, availability of resources at local football fields and demands of the project, the following are the parameters studied and compared for all the setups. Table 3.1 shows the complete analysis.

- **Camera cost:** Thermal camera that we are using is expensive. Installing three thermal cameras together in TC adds up the cost. Wide-angle cameras, on the other hand, are a bit cheaper solution. Another plus in wide-angle camera installation is the less number of camera requirements due to their wide coverage area. This also reduces the total cost of installation.
- **Recognition visibility:** Both thermal and fisheye setups can not provide the solution to person recognition. The fisheye camera has a terrible resolution as the object moves away from the centre. The nearest player to be captured in the field is almost 10m away from the camera, with a resolution of 35x35 pixels. This makes Recognition almost

(a) (b) (c)

**Fig. 3.10:** Example images captured by Thermal camera. **(a)**, **(b)** and **(c)** are from camera 1, 2 and 3 from TC respectively.

impossible. Contrary to that, the nearest person in a wide-angle action camera is of resolution 126x206, which may allow one to recognize a person through the image. A thermal camera with a resolution of 640x480 and a nearest-person resolution of 24x60 is also unsuitable for solving the recognition problem. For every camera setup, the visibility of players decreases as we move towards the outer corner of the field.

- **Day vision:** The observation for the one-day video has shown that every camera setup is performing well enough in daylight conditions. The video from WC is particularly good because of its high resolution.

- **Night vision:** We have observed that thermal cameras are perfect for observing night vision. So TC and RTC perform well in the nighttime. WC also perform well enough except at the outer corner of the field due to low light conditions. The vision in FC becomes unreadable at night, combining low light and lousy corner resolution. We have hypothesized that F-T-C can perform well by making a learning algorithm to learn from thermal feed and applying that to fisheye feed.

- **Occlusion:** The problem of occlusion is easy to handle in colour camera setups than in thermal camera setups. Especially as the objects of interest move farther away from the camera, for catering for the problem of occlusion, we have again hypothesized that learning from the fisheye feed and applying it to the thermal feed can improve to detection of maximum players in F-T-C.

- **Complexity:** Here, we are discussing installation and data preparation complexity. Installation complexity also includes setup preparation. As we increase the number of cameras, the complexity increases with it. Installing FC or RTC is the simplest solution, as they only have one camera. TC and WC are almost equally complex in terms of installation and data processing. For both setups, the cameras' data need to be synchronized. WC needs extra care and build-up while installing as the cameras are tiny in size, and keeping those stable in the northern winds is not easy. F-T-C is equally complex in installation as TC, but data

preparation is relatively complex. The data needs to be synchronized both in time and space. Time synchronization takes more time when both cameras operate at different rates and save the videos in variable clip sizes. It becomes reasonably challenging to process real-time feed.

- **Weather effects:** For the one day of video, thermal setups, i.e. TC and RTC, behave particularly well in all weather conditions. For F-T-C, we have hypothesized that a thermal camera can be used to cater to the weather-related effect in the fisheye setup. Unfortunately, the fisheye setup records weather effects badly as the camera suffers from distortion. On the other hand, even though WC produces very good resolution and quality feed, It still gets significantly affected by the weather. With good person representation in WC, every other environmental effect is visible in the feed.

- **Area coverage:** Out of all setups, FC and F-T-C provide the full field coverage due to the presence of fisheye cameras. TC and WC miss some of the fields below the camera setups. RTC, on the other hand, lacks time-to-time coverage (see, Fig. 3.6a. A significant patch of field can be missed for observation for a particular period. Full-time coverage of the field is a crucial consideration in applying occupancy analysis. All four sides of the fields should be observed in all time slots to consider any player entering or leaving the field. Our application differs from other sports-related computer vision areas where some people play in a sports field for a particular time as a team. The application for such a play could be player detection, player tracking, game analysis and many more. However, this thesis is more focused on occupancy analysis of the field, which can be occupied by people not playing but just stretching in the sun. They can enter and leave the field at any instance of time.

- **Power supply:** Wide angle action cameras operate on battery while all the other setups work on power over Ethernet. Working with cameras with battery-driven options for long recording periods is not convenient. Power over Ethernet and external power is a better option.

- **Privacy preservation:** an essential requirement in our setup is privacy preservation. The data we are dealing with is from local soccer field data. Not everyone likes to be observed while playing at sports fields. It eventually decreases the number of people coming to the fields. Plus, it raises the question of taking the consent of the field users before recording, which could make the whole process complex. Our first and foremost consideration is to make the whole system so that it does not violate the personal privacy of the users of the sports facility.

**Table 3.1:** comparison for five camera setups in terms of cost, recognition visibility, day vision, night vision, occlusion, complexity (includes setup installation complexity and data preparation and handling complexity), weather effects (Shadow, Wind and rain) and Area coverage. In the table,  denotes the worst of all  represents moderate, while  denotes the best or very effective solution. BD stands for battery-driven, while POE stands for power over Ethernet.

| Camera setup | Cost | Rec. Visibility | Day vision | Night vision | Occl--usion | compl--exity | Rain effect | Shadow effect | Wind effect | Area coverage | Power Supply | Privacy preser--vation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FC |  |  |  |  |  |  |  |  |  |  | POE |  |
| TC |  |  |  |  |  |  |  |  |  |  | POE |  |
| WC |  |  |  |  |  |  |  |  |  |  | BD |  |
| F-T-C |  |  |  |  |  |  |  |  |  |  | POE |  |
| RTC |  |  |  |  |  |  |  |  |  |  | POE |  |

The initial analysis gave us know-how on what kind of data we will deal with if we deploy and experiment with it further. Our main requirement from the setup for the particular application of occupancy analysis is full-time and full-area coverage. WC and RTC are deducted from the list for further studies. WC is a battery-driven setup and can hinder long recording time, while RTC can not provide full-field coverage for all-time instances. Based on their complexity, other camera setups will be individually studied in the next chapters.

# 6 Conclusion

This chapter focuses on the preliminary study performed to choose camera setups for further analysis. Different camera setups, their limitations and their analytical performance are discussed in detail. A test setup was employed on the field. Video of a duration of ten hours for all the setups was observed critically. Out of five, three setups, FC, TC, and F-T-C, are chosen for further analysis. The following chapters in the thesis further experiment on above mentioned three setups to study their behaviour in terms of different performance measures.

# References

[1] T. B. Moeslund, *Introduction to video and image processing: Building real systems and applications*. Springer London Ltd, 2012.

[2] A. Tbakhi and S. Amr, "Ibn al-haytham : Father of modern optics," *Annals of Saudi Medicine*, vol. 27, pp. 464–467, 12 2007.

[3] L. Masoner, "A brief history of photography and the camera," https://www.thesprucecrafts.com/brief-history-of-photography-2688527, accessed: 2020-09-20.

[4] R. Kingslake, *A History of the Photographic Lens*, 1st ed. Saint Louis: Elsevier Science, 10.

[5] https://gopro.com/en/us/about-us.

[6] W. R., "Fish-eye views, and vision under water," *Philos Mag 6*, vol. 68, no. 12, pp. 159–162, 1906.

[7] H. Horenstein, *Black and white photography: a basic manual*. Little, Brown and Company, 1983.

[8] R. Kingslake, *Reversed Telephoto Lenses: II. The Fish-Eye Lens*. Academic Press, 2005, p. 145–150.

[9] J. Courbon, Y. Mezouar, L. Eckt, and P. Martinet, "A generic fisheye camera model for robotic applications," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 1683–1688.

[10] C. Eising, J. Horgan, and S. K. Yogamani, "Near-field sensing architecture for low-speed vehicle automation using a surround-view fisheye camera system," *CoRR*, vol. abs/2103.17001, 2021. [Online]. Available: https://arxiv.org/abs/2103.17001

[11] C. Gibson, *Nimrod's Genesis: RAF maritime patrol projects and weapons since 1945*. Hikoki, 2015.

[12] K. J. Havens and E. J. Sharp, "Chapter 10 - thermal imaging applications and experiments," in *Thermal Imaging Techniques to Survey and Monitor Animals in the Wild*, K. J. Havens and E. J. Sharp, Eds. Boston: Academic Press, 2016, pp. 171–244. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128033845000105

[13] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine Vision and Applications, Springer*, vol. 25, no. 1, pp. 145–262, 2014.

# Part III

# Fisheye Camera Setup

# Chapter 4

# Occupancy Analysis of Soccer Fields Using Wide-Angle Lens

Noor Ul Huda, Bolette Dybkjær Hansen, Rikke Gade, Thomas B. Moeslund

The paper has been published in the
*I*nternational Conference on Signal-Image Technology & Internet-Based Systems (SITIS)

*The layout has been revised.*

# Abstract

*Occupancy analysis is critical for resource assessment. This paper presents a novel solution for occupancy analysis in soccer fields, which is needed to assist the management for resource assessment. The analysis is based on player detection in the soccer fields. The process of detection is performed by using one static fish-eye camera, which is achieved by enhancement of the players based on their structural properties using 2-D Gabor wavelet combined with background subtraction. Moreover, the gray scale intensity matching is performed for catering luminance issues. Occlusion is handled through a color and compactness based analysis of expected player regions. It has been shown through experiments that the developed method results in precise analysis of occupancy i.e. an average error of .0094 % for no player in the field, 2.67% for full activity in the field and 3.64% during transition.*


# 1 Introduction

With increasing focus on optimization of resources and time of community, occupancy analysis is becoming vital for the assessment of resources. This technology has been applied to various areas of life e.g. indoor sports arena, in offices for managing light resources, parking areas monitoring and many more [11, 15, 19]. In the field of outdoor sports arenas, specifically soccer fields, authorities are dealing with investing many resources on installing artificial grass without having the knowledge of its significance and utilization. These facilities are highly demanded by local clubs, but very expensive to build, and manual observations are cumbersome and costly. The need of occupancy analysis thus become very important. Hence, there is a need for a system that is cheap and utilizes minimum resources for monitoring sports arenas.

Several solutions [20, 22, 26] have been proposed for sports analysis. Those systems typically employ a large number of cameras, making the whole system expensive. In addition, a general public trust of large-scale camera installations in public facilities are harder to come by due to privacy issues. We therefore apply a one fish-eye camera solution, which is able to capture the complete sports field from distant location. One of the advantage of using fish-eye cameras is the cost effectiveness of this technology. The camera is placed at an aerial view to capture the complete field. This makes people identification difficult, thereby maintaining the privacy. On the other hand, this also makes it hard to detect correct occupancy in the field. As people often appear to be small, they can be misclassified as any other discontinuity in the image. Other than that, because of distant field of view, the challenges of occlusions appear to be more challenging in these images, see figure 4.1.

The novelty of the proposed work is a reliable method for occupancy analy-

**Fig. 4.1:** Scenes with blur and occluded players

sis using fish-eye camera videos. As there is no data set available for fish eye camera soccer videos, so our work is focused on more challenging real-life videos. We utilized the data collected from local soccer field by using only one camera. The focus of the work is to identify the occupancy level in those fields.

## 2 Literature review

Two type of approaches are currently used to analyze occupancy; non visual and image-based approach. In non-visual approach sensor data is used to analyze level of occupancy in the closed room environment. Zhao et. al. [3] performed occupancy analysis in offices. They used Bayesian belief network by utilizing system information e.g. Wi-Fi, chair sensor, GPS location, and keyboard and mouse sensor. Data from passive infrared (PIR) sensors is most commonly used sensor data [4–6]. Energy efficient operation for lighting mainly use the PIR sensors. Carbon dioxide (CO2) level in rooms is also an attractive indicator for analysis, as it is the most obvious consequence of human presence. This indicator is also independent of human movement to some extent [7, 8]. Another way to solve the occupancy problem is to build statistical models for the data [7, 9, 10]. Pedersen et. al. [11] applies image information by defining some rules for trajectory of sensor data to detect occupancy.
The most common disadvantage of existing non-visual methods for occupancy detection is that these sensors can only be utilized in buildings and closed rooms and are not suited for large outdoor fields.
Image based methods on the other hand [12–14] use the video sequences and detect the people to encounter the problem. Zhang et. al. [14] utilized the depth-frame data to detect and track moving people. They used kinetic camera, placed just 4.0 m above the people. Pedersen et. al. [11] follow the same method and detect and track people 2.3-3.0 m below the camera.
Occupancy analysis in sports arenas is previously performed by Gade et. al. [15] and [16]. They used thermal cameras for occupancy analysis of in-

door sports arenas.
All the above mentioned methods either deal with the occupancy detection in indoor buildings or arenas in closed environment. For outdoor environments, the analysis is also performed to monitor occupancy of parking lots [17–19].
Our work is the extension of [15] and [16]. The work in [15] and [16] utilized a set of thermal cameras for occupancy analysis in indoor sports arenas. Here we are using fish-eye camera for capturing outdoor soccer fields. The main part of camera based systems for occupancy detection is detection of people. In case of sports arenas it is the detection of players. A lot of work has been performed on the detection of players in soccer fields by using broadcast videos or more than one camera approach.
A system based on motional graphic feature is proposed by Liu et. al. [20]. They performed the motion analysis and action recognition of players in broadcast videos. Their approach is based on SVM and analysis of optical flow. Another motion detection based system is proposed by Mahmoudi et. al. [21]. The system utilized optical flow analysis using Lucas-Kanade algorithm for motion tracking. Liu et. al. [22] performed player detection and tracking based on Markov Chain Monte Carlo data association using Kalman filter. They used broadcasting video for the test of their algorithm. Khan et. al. [23] proposed a colour based segmentation method and Kalman filter for dealing with occlusion problems while tracking. Hayet et. al. [24] suggested a solution based on point distribution model. They performed tracking by matching model points with a set of similar feature points. They dealt with partial occlusion situations on specific video streams captured through multiple cameras with variable zoom and rotations. Iwase and Saito [25] proposed a solution based on 8 cameras for dealing with occlusion. Beetz et. al. [26] used ontology models of game with motion trajectories in in-camera view of broadcasting videos. Area of interest was first separated by intensity variance and the player identification and tracking was performed by color base segmentation and Blackwellized Resampling particle Filter. Huang et. al. [27] performed players detection by using forward shape analysis-based approach obtained by a trained color histogram-based playfield detector and connected component analysis. They employed Euclidean distance transform to extract skeletons for every foreground blob, and then perform shape analysis to remove false positive detection. Yang et. al. [28] detected soccer players by edge detection combined with Otsu algorithm. They used broadcast videos for testing their algorithm. Another method for detection of players in broadcast videos is performed by Mohammad et. al. [29]. They used Multilayer Perceptron Neural Network as classifier for players classification. Gerke et. al. [31] utilized color histogram and spatiograms in an unsupervised manner for detection. They also tested their algorithm on broadcast videos. Sermetcan et. al. [32] evaluated target players likelihood of being player or not by

using combined appearance and motion model. They used two cameras for deteting and tracking players. M. Manafifard et. al. [33] proposed a detector, that utilizes two-step blob detection (grass-based blob detection followed by an edge-based blob detection) combined with particle swarm optimization (PSO) by assigning sub-swarms to each detected blob.
Our work is based on the occupancy analysis of soccer fields by counting the number of players in the field, using only one wide angle camera to keep the setup simple and cheap. Our detection is based on enhancement of players by using wavelet transform and then color and edge information.

# 3 Methodology

## 3.1 Approach

Occupancy analysis is broadly divided into two categories, i.e. player detection, and occupancy monitoring. The system for the analysis of occupancy uses RGB frames of video sequence as an input to find players in the field, whereas in monitoring, the system detects the changes in the number of players at different time moments. This paper presents a method to perform these analysis. The complete flow diagram of the proposed system is shown in fig 4.2.
The proposed system is composed of following three stages.

- Candidate player region extraction.
- Region classification.
- Players monitoring.

In the first phase, image enhancement is performed to enhance texture of all possible candidate regions for players. Moreover, background subtraction is implemented to remove all candidate enhanced region that belongs to the background i.e. grass, field lines etc. In stage 2, Candidate regions are classified on the basis of color, shape and size. Finally, the number of players are calculated over a large period to determine the level of occupancy.

## 3.2 Data Collection

The fish-eye camera was placed in one public soccer field in Aarhus, Denmark. The soccer field was monitored for four days from 19-22 September, 2016. It captured the whole arena from an aerial view (see figure 4.3). Afterwards, the sequences of videos were transferred to a computer in order to perform the analysis.

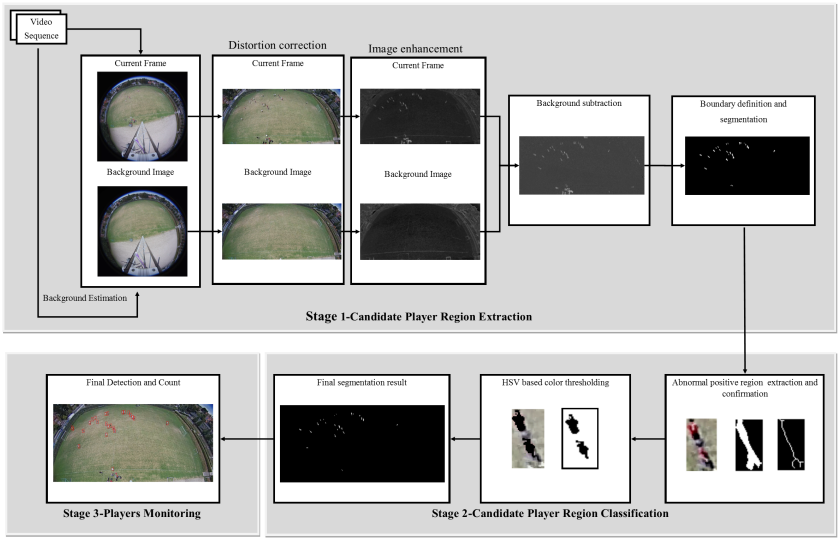



**Fig. 4.2:** Flow diagram for proposed system

## 3.3 Distortion correction

In our system, since the camera films a large area, the lens cause a radial distortion resulting in a convex and round appearance of the image. This distortion must be eliminated, So that the wrapped pixels can be placed at their correct location. The success criteria of most of the distortion correction techniques is the straightening of the curved lines that appears in the image. We, in our work, focused on an image that is clear enough to identify players rather than line straightening. According to Smita's Distortion model [1], the relation between wrapped pixels $q_d = (x_d, y_d)$ and their correct location in the image $q_u = (x_u, y_u)$ is expressed as

$$q_u = q_d(1 + kq_d^2) \tag{4.1}$$

Here $k$ is the lens specific distortion parameter. The above equation can be written in terms of radial dependent magnification $M = 1/(1 + M^2 k q_u^2)$ as

$$x_d = x_u M(k, q_u^2) \tag{4.2}$$

$$y_d = y_u M(k, q_u^2) \tag{4.3}$$

Here $q_u = \sqrt{(x_u^2 + y_u^2)}$, so it is suggested to have the model in terms of $q_u^2$. The points from the distorted image are mapped on the model image on corrected position $q_u = (x_u, y_u)$ in order to magnify and perceive a better understanding of player's positions in our region of interest. The final image does not correct the straight lines but produce a better, magnified and clear image necessary for occupancy analysis.

**Fig. 4.3:** Frame acquired by the fish-eye camera

## 3.4 Player enhancement

As this work takes into account the real scenario with natural grass and only one camera view, It is insufficient to use only background subtraction and color based segmentation to give some satisfactory results. Pixels belonging to one single player may appear as separate blobs while using only background subtraction. It also results in missing the detection of many players, as the size of the players is small in the image. Therefore, we are using pixel based enhancement and segmentation in our work.
Our method enhance the soccer players in the model image using Gabor wavelets. Gabor wavelets are flexible in terms of different frequencies and orientations, which is useful for enhancing players based on their textures. This emphasizes on locating player in the soccer field as single blobs regardless of illumination conditions. Wavelets perform as low-level oriented edge discriminators [2]. Since the players have directional patterns, 2-D Continuous Wavelet Transform are the best option to use for enhancement due to their directional selectiveness capability. They can be used to detect even slightly slanted features by using different frequency tunings. These wavelets are defined as

$$\psi_G(x) = exp(jk_o)exp(-\frac{1}{2}|Ax|^2) \tag{4.4}$$

(a) (b)

**Fig. 4.4:** a) Image after distortion correction b) Image after applying wavelets

Here A ia elongation of the filter defined by $\begin{bmatrix} \varepsilon^{-\frac{1}{2}} & 0 \\ 0 & 1 \end{bmatrix}$ and $k_o \in R$ is a vector that defines the frequency of the complex exponential. The enhancement equation can be defined as

$$T_\psi(b,\theta,a) = C_\psi^{-\frac{1}{2}} a \int exp(jkb)\hat{\psi}^*(ar_{-\theta}k)\hat{I}(k)d^2k \tag{4.5}$$

Here $a$ and $b$ are the dilation parameter and displacement vectors respectively, $\theta$ defines the angle of rotation, $C_\psi$ is the normalization constant for the particular wavelet $\psi$ and $I$ is the image representation as square integral.
For each pixel position in an image the $max(T_\psi(b,\theta,a))$ is computed for spanning $\theta$ from $0^o$ to $170^o$ at steps of $8^o$. Inverse green image plane is used for enhancement as it gives best contrast of players over nearly green grass. Transformed model image and the image after enhancement is shown in the figure 4.4.

## 3.5 Backgrond subtraction and player detection

The enhanced soccer field image contains some lines and noise of the field which may cause the separation of the players from the field difficult. To decrease this noise we have integrated our idea with a very useful and convenient background subtraction approach.
Background subtraction is the well known approach for motion detection. It is based on assessment of current frame by comparing it with a reference frame. Let the enhanced reference background model obtained after using the Guassian Mixture model [30] be $B_E(x,y)$ and the enhanced current frame be $I_E(x,y)$, then a pixel $(x,y)$ is supposed to be a foreground pixel if it differs from the background model $B_E(x,y)$ more than twice the standard deviation [34]. Gray level pixel intensity matching with the reference background frame is applied on the images before enhancement and subtraction to minimize the changing illuminance effects. Figure 4.5 shows the resultaning image obtained after background elimination.

The implemented 2D wavelet based background subtraction model use only the edge and energy information of each point in a frame. Each point in the image is evaluated in a small sliding window to be categorized as moving or static. The small edges that may be discontinuities in natural grass are not included in the final image. This leads to the final image that contains the players only.

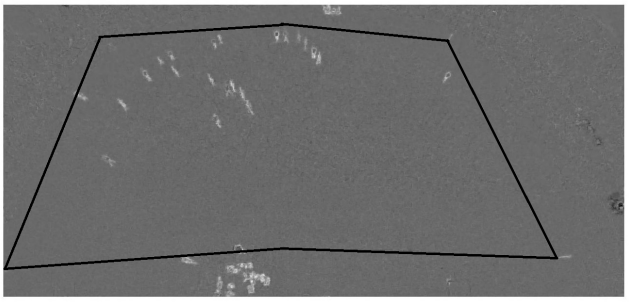


**Fig. 4.5:** Final response after background elimination shown with the applied definition of boundary

As seen in fig. 4.5 detection of players in the field is a difficult task with one wide angle lens as the size of a player in the image are less 30 pixel height. This often results in unclear image content. Shapes and colors seems to be blurred and share the same boundary.
As we are using static camera configurations, so we can define the boundary of the field ourself, as natural does not always give good contrast. Region of interest boundaries are defined and threshold on the intensities in the enhanced d image is applied to extract the candidate player regions. Morphology is used to remove small noise pixels. Enhancement of the frame helps us in extracting the candidate players as full-connected blobs but further processing is required to separate the falsely connected candidate regions (occlusions).
Candidate player regions that may contain occlusions are grouped into abnormal regions. Abnormal regions are separated based on expected length of major axis, minor axis and finally the compactness of the pixels in those specific regions. Furthermore, the area of the skeletonized abnormal region is calculated to confirm the occlusion. In order to split the abnormal regions, color constraint are applied in the regions. The color model for the shirts of the players is defined in HSV space and the value of each pixel $(x, y)$ in candidate region is compared with the predefined color model for all the abnormal regions. If the pixel in that candidate region lie in the model space it is labeled as true shirt pixel otherwise it is discarded. This means that only the points belonging to the upper body of the player are labeled, as true pos-

itive (for that particular abnormal region), and all other pixels are labeled as negative. This will result in separating the blobs for each player by eliminating parts of body other than the upper body. This is shown in fig. 4.6. Final players detection is shown in the fig. 4.7.

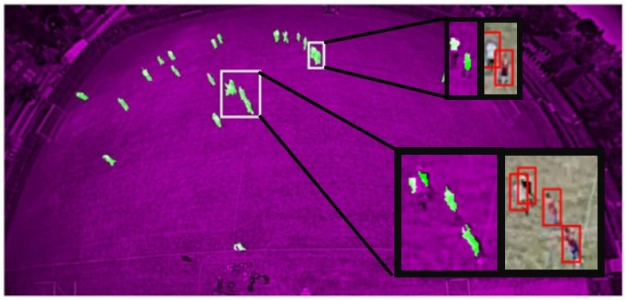


**Fig. 4.6:** Occluded players after applying color and compactness constraints

**Fig. 4.7:** Finally detected players in soccer field

# 4 Experimental results

As there is no publically available data set for fish-eye camera videos of soccer fields, so experiment is carried out placing our own camera in the soccer field in Aarhus, Denmark. A static Hikvision Fish-eye network camera, with a resolution of $1280 \times 1280$ pixels and a field-of-view of $360^o$, is used to cover the experiment. It captures the video at 10 frames per second and covers the complete field areas and some surroundings. 6000 frames of videos are used to test the proposed method. The ground truth is calculated by manually counting the number of persons, frame by frame, in the video. Player's occupancy i.e the number of players in the field over the time, shown in figure 4.8, is computed by taking the mean for the window of 100 samples. It can

be visualized that the proposed scheme detects the players occupancy with low error rate. It is observed that the error is very low when detecting empty fields, an average error of 0.0094 %. With full activity on the field (22-24 persons) the average error is found to be 2.67%, while during the transition, when lot of players leave in or out of the field, the error is 3.64%.

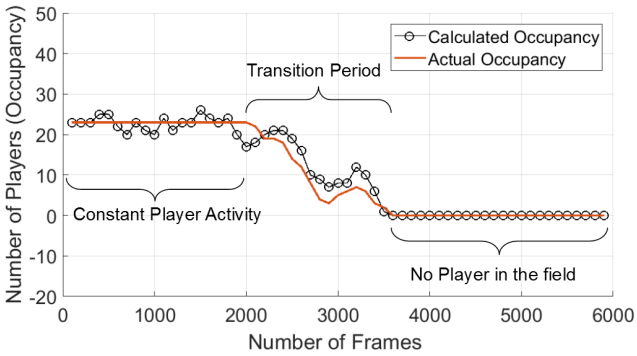



**Fig. 4.8:** Occupancy analysis over 6000 samples

# 5 Conclusion

# References

[1] S. L. Darvatkar, and S. U. Bhandari *Implementation of Barrel Distortion Correction Algorithm for Wide Angle Camera Based Systems*, i-manager's Journal on Image Processing, vol. 4, no. 1,pp. 43-48. 2017.

[2] J. P. Antoine, P. Carette, R. Murenzi, and B. Piette, *Image analysis with two-dimensional continuous wavelet transform*. Signal Process, vol. 31, pp. 241–272, 1993.

[3] Y. Zhao, C. Lei, and J. C. Patterson, *A PIV measurement of the natural transition of a natural convection boundary layer*, .Experiments in Fluids, Springer, January 2015.

[4] Varick, L. Erickson, M. A. Carreira-Perpiñán, and A. E. Cerpa, *Occupancy Modeling and Prediction for Building Energy Management*, ACM Transactions on Sensor Networks (TOSN), Springer, vol. 10, no. 3, April 2014.

[5] R. J. Dobbs, and M. H. Brandon, *Model predictive HVAC control with online occupancy model*, Energy and Buildings, Elsevier, vol. 82, pp. 675-684, October 2014.

[6] B. Dong, B. Andrews, K. P. Lam, M. Höynck, R. Zhang, Y. S. Chiou, and D. Benitez, *An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network*, Energy and Buildings, Elsevier BV, vol. 42, pp. 1038-1046, 2010.

[7] C. Jiang, M. K. Masood, Y. C. Soh, and H. Li, *Indoor occupancy estimation from carbon dioxide concentration*, Energy and Buildings, Elsevier, vol. 131, pp. 132-141, November 2016.

[8] H. Han, K. J. Jang, C. Han, and J. Lee, *Occupancy estimation based on carbon dioxide concentration using dynamic neural netork model*, Proceedings of the 34th AIVC - 1st venticool Conference , pp. 25-26, September 2013.

[9] B. Dong, and K. P. Lam, *Building energy and comfort management through occupant behaviour pattern detection based on a large-scale environmental sensor network*, Journal of Building Performance Simulation, vol. 4, pp. 359-369, July 2011.

[10] X. Lianga, T. Hong, and G. Q. Shena *Occupancy data analytics and prediction: A case study*, Building and Environment, Elsevier, vol. 102, pp. 179–192, June 2016.

[11] T. H. Pedersen, K. U. Nielsen, and S. Petersen, *Method for room occupancy detection based on trajectory of indoor climate sensor data*, Building and Environment, Elsevier, vol. 115, pp. 147-156, April 2017.

[12] Y. Benezeth, H. Laurent, B. Emile, and C. Rosenberger *Towards a sensor for detecting human presence and activity*, Elsevier International journal on Energy and Buildings, 2011, 43, pp. 305-314.

[13] H. C. Shih, *A robust occupancy detection and tracking algorithm for the automatic monitoring and commissioning of a building*, Elsevier International journal on Energy and Buildings, 2014, vol. 77, pp. 270-280.

[14] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li, *Water Filling: Unsupervised People Counting via Vertical Kinect Sensor*, IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), 2012, Beijing, China.

[15] R. Gade, A. Jørgensen, and T. B. Moeslund *Long-Term Occupancy Analysis Using Graph-Based Optimisation in Thermal Imagery*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR),, pp. 3698-3705, 2013.

[16] R. Gade, A. Jørgensen, and T. B. Moeslund, *Occupancy Analysis of Sports Arenas Using Thermal Imaging.* In Proceedings of the International Conference on Computer Vision Theory and Applications, pp. 277-283, 2012.

[17] C. G. Postigo, J. Torres, and J. M. Menéndez *Vacant parking area estimation through background subtraction and transience map analysis*, IET Intelligent Transport Systems, vol. 9, no. 9, pp. 835-841, 2015.

[18] B. Karunamoorthy, R. SureshKumar, and N. JayaSudha, *Design and Implementation of an Intelligent Parking Management System using Image Processing*, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 4, no. 1, January 2015.

[19] G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, and C. Vairo, *Deep learning for decentralized parking lot occupancy detection*, Elsevier journal of Expert Systems with Applications, vol. 72, pp. 327-334, April 2017.

[20] G. Liu, D. Zhang, and H. Li, *Research on Action Recognition of Player in Broadcast Sports Video*, International Journal of Multimedia and Ubiquitous Engineering [SCOPUS], vol. 9, no. 10, pp. 297-306, 2014.

[21] SA. Mahmoudi, M. Kierzynka, P. Manneback, and K. Kurowski, *Real-time motion tracking using optical flow on multiple GPUs*, Bulletin of the Polish academy of sciences, Technical Sciences, vol. 62, no. 1, 2014.

[22] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang and H. Wang, *Automatic player detection, labeling and tracking in broadcast soccer video*, Pattern Recognition Letters, Elsevier Science Inc., vol. 30, nr. 2, pp. 103-113, 2009.

[23] MM. Khan, TW. Awan, I. Kim and Y. Soh, *Tracking Occulded Objects Using Kalman Filter and Color Information*, International Journal of Computer Theory and Engineering, vol. 6, no. 5, 2014.

[24] J. B. Hayet, T. Mathes, J. Czyz, J. Piater, J. Verly and B. Macq, *A modular multi-camera framework for team sports tracking*, Advanced Video and Signal Based Surveillance, IEEE 2005.

[25] S. Iwase and H. Saito, *Tracking soccer players based on homography among multiple views*, Visual Communication and Image Processing, vol. 5150, pp. 283-292, 2003.

[26] M. Beetz, S. Gedikli, J. Bandouch, B. Kirchlechner, N. V. Hoyningen-Huene and A. C. Perzylo, *Visually Tracking Football Games Based on TV Broadcasts*, 20th international joint conference on Artificial inteligence, 2007.

[27] Y. Huang, J. Llach, and S. Bhagavathy, *Players and Ball Detection in Soccer Videos Based on Color Segmentation and Shape Analysis*, MCAM Springer ,pp. 416–425, 2007.

[28] Y. Yang and D. Li, *Robust player detection and tracking in broadcast soccer video based on enhanced particle filter*, Journal of Visual Communication and Image Representation, March 2017.

[29] A. H. Mohammad, *An MLP-Based Player Detection and Tracking in Broadcast Soccer Video*, International Conference on Robotics and Artificial Intelligence (ICRAI), 2012, Rawalpindi, Pakistan.

[30] P. Kaewtrakulpong, R. Bowden, *An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection*, In Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS01, VIDEO BASED SURVEILLANCE SYSTEMS: Computer Vision and Distributed Processing, September 2001.

[31] S. Gerke, S. Singh, A. Linnemann, and P. Ndjiki-Nya, *Unsupervised color classifier training for soccer player detection*, Visual Communications and Image Processing (VCIP), 2013, Kuching, Malaysia.

[32] S. Baysal, and P. Duygulu, *Sentioscope: A Soccer Player Tracking System Using Model Field Particles*, IEEE Transactions on Circuits and Systems for Video Technology, vol.: 26, no.: 7, July 2016.

[33] M. Manafifard, H. Ebadih, and A. Moghaddam, *Multi-player detection in soccer broadcast videos using a blob-guided particle swarm optimization method*, Multimedia Tools and Application, vol. 76, no. 10, pp. 12251–12280, May 2017.

[34] P. Spagnolo, N.Mosca, M.Nitti, and N.Distante, *An Unsupervised Approach for Segmentation and Clustering of Soccer Players*, Coference Proceeding in International Machine vision and Image Processing, 5-7 Sept. 2007, Kildare, Ireland

# Part IV

# Thermal Camera Setup

# Chapter 5

# Estimating the Number of Soccer Players using Simulation-based Occlusion Handling

Noor Ul Huda, Kasper Halkjær Jensen, Rikke Gade, Thomas B. Moeslund

The paper has been published in the
Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

*The layout has been revised.*

# Abstract

*Estimating the number of soccer players is crucial information for occupancy analysis and other monitoring activities in sports analysis. It depends on player detection in the field that should be independent of the environment and light conditions. Thermal cameras are therefore a better option over normal RGB cameras. Detection of non-occluded players is doable but precise estimation of number of the players in groups is hard to achieve. Here we propose a novel method for estimating number of the players in groups using computer graphics and virtual simulations. Occlusion conditions are first classified by using distinctive set of features trained by a bagged tree classifier. Estimation of the number of players is then performed by maximum likelihood of probability density based approach to further classify the occluded players. The results show that the implemented strategy is capable of providing precise results even during occlusion conditions.*


# 1 Introduction

Soccer is the most popular sport around the world [34]. The application of soccer video analysis includes strategy understanding, player action recognition, occupancy analysis and many more. Estimation of number of players is the foundation of understanding soccer especially if we need to know the occupancy in a particular field. The occupancy analysis can be achieved by counting the number of players with respect to the time stamp over a large period [11]. Player detection and correct estimation of a number of players in the sports field is the basic step in every sports analysis. A number of solutions [36] have been proposed for soccer analysis. These solutions normally employ a large number of cameras or used broadcast videos for analysis. This makes the whole system very complex to deploy in local sports fields. Furthermore, the communal trust of large-scale, high-resolution camera systems in public fields is harder to come by due to the general privacy issues.

Precisely estimating the number of players is a challenging topic due to various factors. These factors consist of occlusion, motion blur, varying illumination, outdoor weather, changing player sizes and inconsistency in appearance of the players. Even though multi-camera solutions improves the precision by providing more information [8]. They also increase the hardware and complexity of the whole system. RGB cameras are also effective in many cases but they are challenged in varying illumination conditions.

Consequently, we propose a thermal camera based solution for estimating the number of players in groups and counting. Three thermal cameras are installed on a single pole. The setup is able to capture the complete soccer field from a distant location. One of the advantages of using three thermal cameras on the same location is its ease of installation. The other obvious advantage

of the thermal camera over normal cameras is the privacy preservation, because thermal view makes people identification almost impossible. Contrary to this, it is hard to detect and estimate the correct number of players, specially when they are in groups. This is because thermal cameras provide less textual information about player appearance. The main contribution of this work is a simulation-based approach that efficiently deals with occlusion in the thermal camera view.

## 2 State of the art

**Player Detection in Soccer Field**

Various solution for the detection of players in soccer fields have been proposed. These methods include background subtraction and spectator region extractions [5, 8, 19, 20, 24, ]. However, there may remain some noisy regions in the image (line segments and other discontinuities in the image). Yoon *et al.* [43] proposed a solution based on player regions separation by defining thresholds for size, compactness, ratio of vertical to horizontal length and color distribution. Haung *et al.* [18] presented a shape-based player detection method in order to remove noisy areas from connected components. Yao *et al.* [42] used the confidence map for segmentation of players in broadcast videos. The confidence map is generated from the output of a Hough forest.

A Motion graphic feature based system is proposed by Liu *et al.* [25]. Their approach is based on motion analysis and action recognition of players using SVM and optical flow analysis in broadcast videos. Mahmoudi *et al.* [28] proposed another motion based system. Their system also utilized optical flow analysis with Lucas-Kanade algorithm for detection and tracking of players. A method based on Markov Chain Monte Carlo data association and Kalman filter is proposed by Liu *et al.* [26]. A combined appearance and motion model for evaluating player regions is proposed by Sermetcan *et al.* [1]. They used two camera system for detection and tracking players. Direkoglu *et al.* [7] proposed an 8 camera based system for detection of player in the field. They proposed a diffusion equation based solution to make the whole system invariant of color and rotation information.

Beetz *et al.* [2] utilized ontology models of game together with motion trajectories of players for detection. They suggested Blackwellized Resampling particle Filter for tracking of players. Intensity variance and color based segmentation is used for player segmentation in their work. Liu et al. [27] proposed a context-conditioned motion based tracking model. They work is based on the fact that the player response in an existing situation in only a limited number of ways. Yang *et al.* [41] proposed edge detection and threshold based player detection. They used broadcast videos for test-

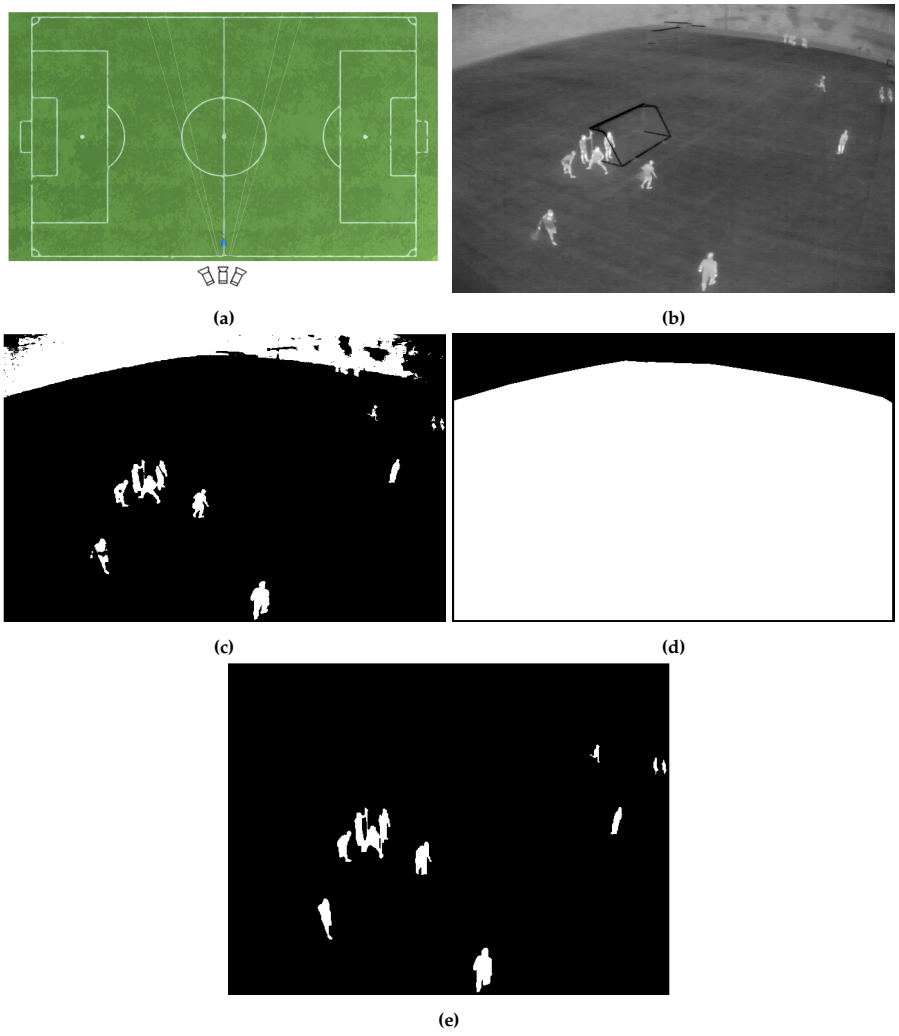



**Fig. 5.1:** (a) shows the top camera view of the soccer field with camera positions, (b) is the reference frame from left camera view, (c) shows the binary Image, (d) is the mask of the image and (e) is the final image we get after applying background mask and morphology.

ing their algorithm. Heydari *et al.* [17] used Multilayer Perceptron Neural Networks for classification of players in broadcast videos. Gerke *et al.* [15] proposed color histogram and spatiograms for the detection of players. They enhanced their work using histogram based features for the identification of players [14]. They also tested their algorithm on broadcast videos. Most of the literature is either based on evaluation on broadcast videos or large camera setup is employed for player detection and tracking. This leads to the lack of more simple and robust approach for estimating number of players.

**Occlusion Handling**

Occlusion is one of the major problems while dealing with sports videos. Khan *et al.* [31] proposed a color based segmentation method and Kalman filter for dealing with occlusion problems during tracking. Hayet *et al.* [16] performed detection based on point distribution model. They dealt with partial occlusion situations on specific video streams captured through multiple cameras with variable zoom and rotations. Iwase and Saito [20] proposed a solution based on 8 cameras for dealing with occlusion. Sabirin *et al.* [33] proposed free viewpoint based approach to cater occlusion while tracking. Kristoffersen *et al.* [23] perform people counting and occlusion handling by using stereo thermal camera setup in the street. They perform 3-D reconstruction and deal with occlusion based on clustering and tracking of the 3D point clouds. Manafifard *et al.* [29] suggested a detector that performs two-step blob detection (grass-based blob detection followed by an edge-based blob detection). They handled occlusion by a blob-guided PSO multi-player detection algorithm. Gade *et al.* [11] proposed a system based on identifying the occlusion in thermal cameras by defining thresholds on length and width of the blobs. They also enhance their work of player counting in indoor and outdoor sports arenas using constraint information of the stable periods by Graph search optimization [13]. In most of the literature, either the occlusion is handled in tracking of the players or complex system is employed to capture the videos at multiple angles. Choosing a threshold is also a compromise between false positives (spurious occlusions) and false negatives (missed occlusions) detections.

**Thermal cameras**

Automatic identification of human body includes both the visual and thermal information [4, 22, 32, 37, 40]. A comprehensive survey regarding thermal cameras and their application is performed by Gade *et al.* [12] Gade *et al.* [10, 11, 13] also proposed player counting and occupancy analysis by using thermal cameras in indoor sports arenas. Other work in the domain of thermal cameras for video analysis and human detection is pedestrian detec-

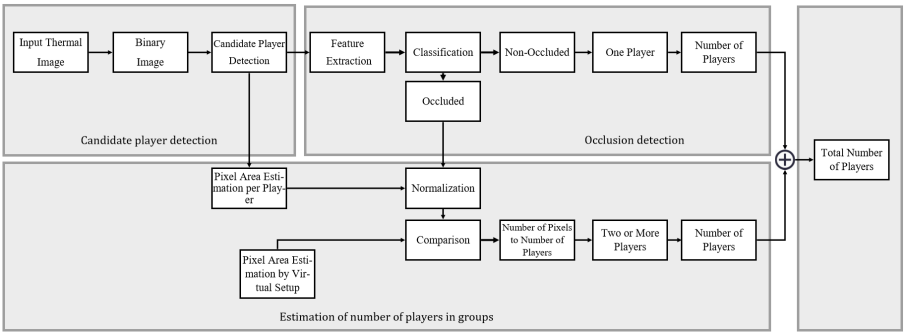



**Fig. 5.2:** Flow diagram for the proposed system.

tion and counting [23, 39]. Most of the work for utilizing thermal cameras in sports analysis is related to indoor sports arenas with a relevantly closed environment and small area of interest. Large outdoor sports fields are yet to be analyzed with thermal camera setups.

# 3 Proposed Method

Most vision systems for detection and counting of players in a sport field are either complex, in terms of number of cameras and controlled light environment, or they lack any supervised algorithm for the detection of player on a soccer field. In this work, we proposed a three staged supervised player detection and counting system i.e. candidate player detection, occlusion detection and estimation of number of players in each occluded group.

Player detection in soccer has always been a challenging task because of various factor i.e. weather (wind, rain, snow, clouds, etc.) and varying light conditions. Moreover, the shape, geometry and size of the player in the field vary with the angle and position of the camera.

To cope with all these issues here we propose a fixed three thermal camera-based solution that is independent of varying light and weather conditions and for the evaluation of our algorithm, we apply our approach to the outdoor field of soccer. The camera setup is shown in figure 9.4a. The proposed approach for estimating the number of players consist of the following steps. Given a video frame by the thermal camera, we compute a binary image. We have assumed that a player is any human on the field that could be a team player or a referee. From the resulting image a feature vector is extracted from each blob and afterwards, a bagged tree classifier is implemented to separate blobs into occluded and non-occluded players. Occluded players are then fed to a maximum likelihood of density estimation analysis to estimate the actual number of players in each occluded blob.

The whole process, of occlusion detection and estimating number of players in the occluded blob, is illustrated in the figure 5.2 and described in details in the following.

## 3.1 Player Detection

The thermal camera captures grey scale where a warm object, which is a player in our case, appears to be brighter than the surroundings and background. The first step in our algorithm is to detect and separate these objects from the image. Maximum entropy based thresholding that finds the threshold value based on the sum of entropies is employed for the segmentation of these light objects [21]. There may remain some blobs and noise outside the field because of intensity variations or spectators. Those blobs are removed from the image by a manually marked geometry based field mask. Blobs smaller than specified minimum area are discarded as they may belong to some noise and morphological closing is applied to join the other small blobs. The blobs are then labeled using a contour-finding algorithm [35].

## 3.2 Occlusion Detection

People standing beside or behind each other often merge into one big blob. These blobs come under the category of occlusion. One of the major contribution in this paper is occlusion handling by utilizing only blob information. Where the aim is to decide if a blob is one player or more than one (occluded player).

Images obtained from a thermal camera do not carry much textual or color information. Also, the players have different posture, shape and size that depending on the distance and orientation of the cameras. Moreover, the players on the border of the field appear too small to detect. Therefore, the normal state of the art feature extraction methods fails. The only thing that can be utilized is the shape and orientation of the blobs.

#### Feature Vector Information

Players, whether occluded or non-occluded, appear larger near the camera. However, at the same pixel position, an occluded blob appears to be larger than a non-occluded blob. The non-occluded blob always has a vertically oriented shape, as players are always vertically oriented from the camera point of view. For an automated system to distinguish between occluded and non-occluded regions, a feature vector based on the blob information is formed for each candidate region. If a binary image $I$ contains $N$ potential candidate regions, then the set representation for an image $I$ is $I = \{I_1, I_2....I_N\}$. Each candidate region is considered as a sample for classification and represented

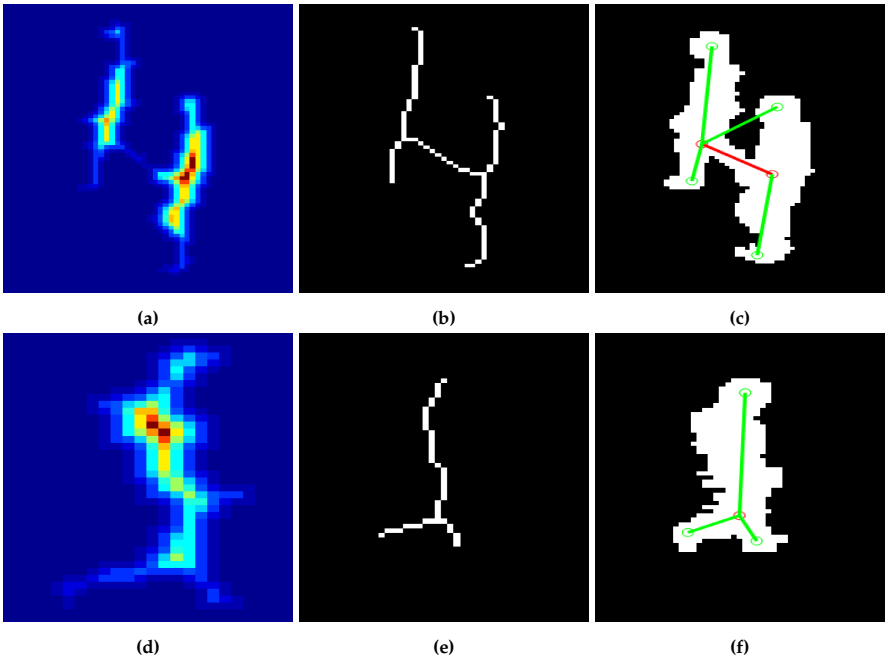



**Fig. 5.3:** (a), (b) and (c) show the heat map, skeleton mask and connected points of an occluded region, (d), (e) and (f) show the heat map, skeleton mask and connected points of a non-occluded region. Note that red circles in (c) and (f) show the connected points whereas green circles and lines are the branch points.

by a feature vector containing all $M$ features, i.e. for a sample non-occluded blob $I_i$ the feature vector is $I_i = \{f_1, f_2, f_3....F_M\}$, where $i = \{1, 2, 3, ...N\}$. The set of features utilized here are:

- **Connected point slope:** The connected points and branch points say $C(x, y)$ of the skeleton of the blobs are founded by using [9]. In the case of non-occluded blobs all the founded points, $C = \{C_1, C_2, C_3, ....C_k\}$, are connected in a single vertical symmetry. While the blob of an occluded group of players normally have two or more vertical symmetries. In these cases, slopes between each connected point provides a useful information to distinguish between occluded and non-occluded players' blobs and is computed according to the following equation,

$$slope_k = \frac{y_{k+1} - y_k}{x_{k+1} - x_k} \tag{5.1}$$

  Where $k$ represents a point in the set of all potential connected points, $K$ from $1 : k$, with $k = 1$ is the top branch point and $K = k$ is the bottom-most branch point. In the case of non-occluded player blobs the slopes

would be greater between each connected point, whereas the in case of occlusion, where there are two or more vertical symmetries the slopes would be smaller.

- **Connected point distance:** This is the distance between the connected points in the skeleton of a blob. It can be observed in figure 5.3d that for some connected points, smaller slopes can occur in non-occluded cases as well. In that cases the distance between the connected points contain useful information. So, If two connected points with smaller slope have a larger distance between them, then they probably belong to an occluded blob otherwise not, as shown in the figure 5.3c.

- **Convex area:** It is the area of the convex hull of a blob. The convex area of occluded blobs appears to be larger than the convex area of non-occluded blob at the same position. But as we move away from the camera both occluded and non-occluded blobs appear to be small. So the blob area with respect to pixel distance from the camera is considered as a feature in our case. The pixel distance is calculated by equation 5.2.

$$y' = \sqrt{| \frac{640}{2} - y |^2 + | 480 - x |^2} \tag{5.2}$$

  Here $x$ and $y$ are the pixel locations in image $I$ (varying from 0 to 640 and 0 to 480 respectively). $y'$ is the pixel distance with respect to the camera (see figure 5.4b) and 640 x 480 is the size of the image.

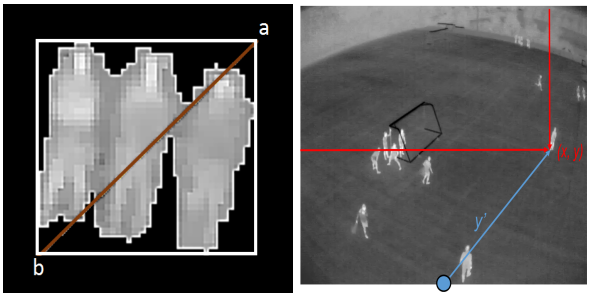
- **Diagonal Length of bounding box:** The diagonal distance of the bounding box (figure 5.4a) is the last feature to be used for classification. This distance would be larger for occluded blobs then non-occluded blobs.

#### Classification using Bragged Tree Classifier

Feature extraction is followed by the implementation of Bragged tree classifier [3] to distinguish between occlusion and non-occlusion blobs. The training dataset used for the classifier includes 1700 non-occluded and 120 occluded player blobs samples collected from over three different soccer videos. Evaluation of the proposed classifier was performed using k-folds cross validation with folds selected to be 5. Results are explained in section 4.2.

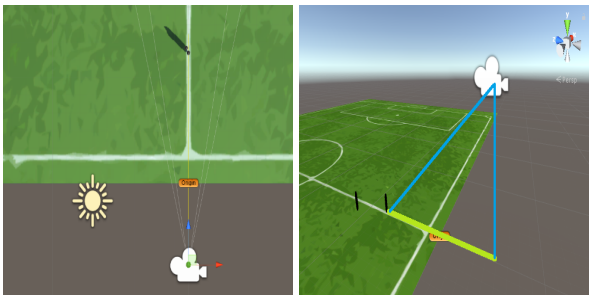
## 3.3 Estimating the number of players

Here we present a novel method for estimating the number of players in an occluded blob. Our method estimates the number of players in a blob by




(a) (b)

**Fig. 5.4:** (a) The diagonal length of a bounding box is measured between the corners 'a' and 'b', (b) shows the pixel distance calculation where the blue lines are the pixel distances and red lines are the original distances in the image plane.



(a) (b)

**Fig. 5.5:** Virtual setup. (a).Top View of our setup, (b). Side View of our setup.

comparing the size of the detected blob with different likelihoods of learnt blob sizes created in virtual environment. A virtual setup of a football field with real world field coordinates, camera height, viewing angle and resolution is created using unity [38]. A human body is modeled as cylindrical blobs. The height, depth and width of the cylinders are taken as standard person height and width. Virtual player occlusion is created by considering the fact that occlusion can occur in a finite number of possible ways. All of these are simulated and a likelihood density is learned for each distance from the camera. Shadows are not considered in our work as the background surface is non-reflecting and no shadow occurs in case of thermal view. The virtual setup is illustrated in figure 5.5

The process of simulating data for occlusion of two persons is

1. One static player is placed at minimum possible distance from the camera in the field.

2. The second player is shifted horizontally towards the first player from

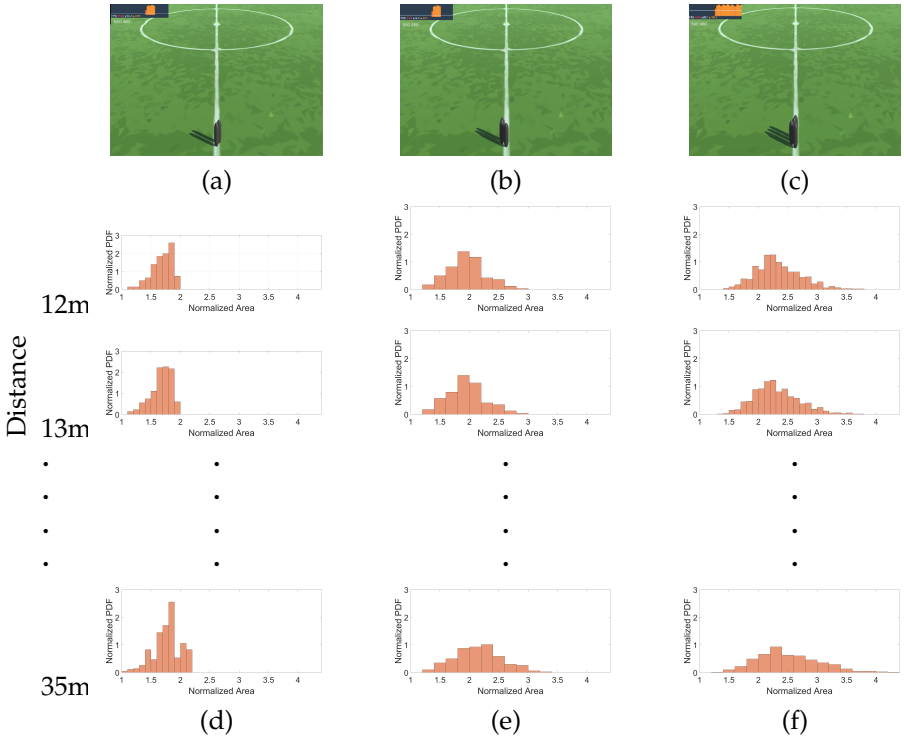



**Fig. 5.6:** (a), (b) and (c) are the examples of virtual projections of two, three and four occluded players, respectively. (d), (e) and (f) are the normalized probability densities with respect to distance.

right to left in steps of 0.05 meter until they occlude in the 2D camera view.

3. The instant occlusion occurs the algorithm measures the combined pixel area of the blob.

4. The second player is shifted until the players are non-occluded in the camera. For each step the pixel area of the blob is measured.

5. The second person now shifts 0.05 meters above the previous position and steps 6.4b-6.4e are repeated until no occlusion is present.

6. The first static person is then moved 1m further away from the initial position and the process (steps 6.4b-6.4f) is repeated for the entire field.

Since the size of a blob to a large degree is independent of the viewing direction, the steps above are only required for one particular viewing direction, and hence the size of a blob only depends on the distance from the camera.

The process for three and four players follows a similar procedure except that it forms more combinations of static and moving players. The process could be repeated for higher number of players but the data set we are using is having a maximum of four occluded players.

**Maximum likelihood based density estimation**

The processes above result in 9978 possible occlusion combinations for two players, 12401 possible occlusion combination for three players and 33001 possible occlusion combination for four players. The size of each combination is normalized by the size of one simulated person at that particular distance. This results in a likelihood distribution that expresses the size as a function of the number of persons. This is illustrated in figure 5.6. After classification of

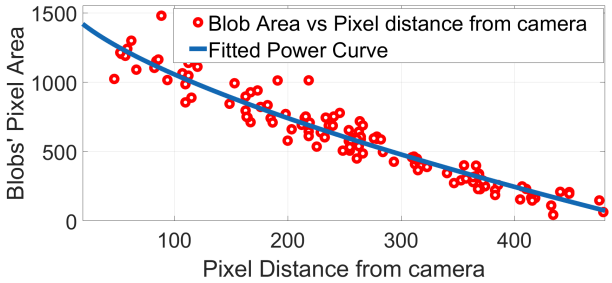



**Fig. 5.7:** Relationship between blob area and camera distance.

occlusion, see figure 5.2, pixel distance with respect to camera is first calculated for each occluded blob (see equation 5.2). Then the blob is normalized with respect to the pixel area of a one-player blob, so the measure can be compared with the simulated data. In order to perform this normalization, an analytic function is learned from one-player blobs and their distances to the camera, see figure 6.3. This lead us to formulate a general relationship between pixel distance from the camera and average pixel area of a person that can be used for estimating the area of one person at every pixel location. The relationship is dependent on external parameters like camera height and tilt angle.
Normalized area with its pixel distance is compared with the virtually generated likelihood distributions. The one that matches best determine the actual number of players in a particular blob. In case, the pixel distance does not match with any of the learn likelihood densities, the nearest neighbor distance is considered in that case.

# 4 Experiments

## 4.1 Data and Setup

As there exist no publicly available soccer data captured with thermal cameras, we have used our own captured data for evaluation purposes. The video is captured with an AXIS Q1922 LWIR sensor with 57 degrees of horizontal Field of View (FOV) and a resolution of 640x480. The camera is placed on a pole that is 4.6 meter away from the field at a height of 10.5 and tilt angle of 27◦. The data that are used for the testing contain 5 minutes of video with 8990 frames containing 71443 players. The ground truth is marked by manually counting the number of players in each frame.

## 4.2 Results

In this paper, we present the results for the left view camera. Here the evaluation of our features with Bagged tree classification model. This evaluation includes the comparison of our features with state of art human detection histogram of oriented gradients (HOG) features [6] which are still used in many state of the art player detection algorithms [1, 15]. [7, 15, 29] have also performed this comparison analysis. HOG is trained on grey scale images for classification of occluded and non-occluded blobs. Figure 5.8 shows the comparison of two features in terms of ROC. Other comparison measures for evaluation of our features are presented in Tables 5.1.

Clearly our proposed features gather with classifier performed better than

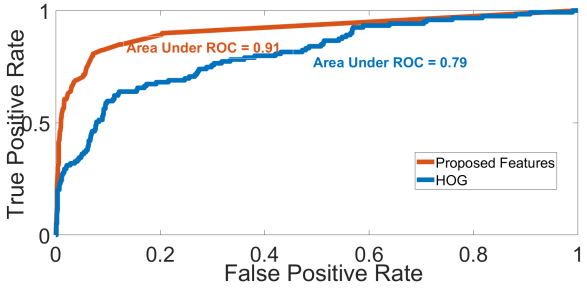
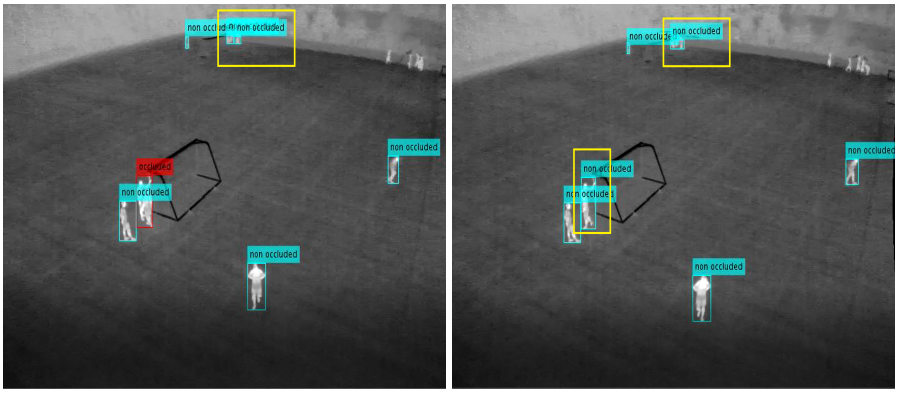



**Fig. 5.8:** Receiver Operating Curve for the [6] and proposed method.

[6]. 100% accuracy can not be achieved due to more false negative cases because of the factors that include fully occluded cases and occlusions in farther blobs interpreted as non-occlusions. This is because the blobs appear to be too small to detect and classify, see figure 5.9.

| | Acc | P % | R% | AUC | TT |
|---|---|---|---|---|---|
| HOG+SVM [6] | 94.3 | 62.1 | 77.3 | 0.79 | 142.0sec |
| Proposes | 96.1 | 77.3 | 84.1 | 0.91 | 1.5sec |

**Table 5.1:** Comparison of the methods for occlusion detection with [6] in terms of accuracy (Acc), precision (P), Recall (R), area under receiver operating curve (AUC) and training time (TT).




**(a)** Frame 400 **(b)** Frame 480

**Fig. 5.9:** Non-Occluded vs Occluded, yellow boxes show some false negative cases.

Next we evaluate our method for estimating the number of players compared with the ground truth. The results are is presented in figure 5.10. Figure 5.11 shows some qualitative results after estimation. The results presented demonstrate the precision of the proposed method. Another important observation from figure 5.10 is that the proposed approach yields better accuracies in former frames. The reason is that most of the players appear at the farther boundary of the field in the last part of the test data. This makes the occlusion detection more challenging and hence estimation become more uncertain.

For the quantitative evaluation, a comparative analysis of our proposed algorithm with previously proposed algorithms is presented in table 5.2.

Our system clearly outperforms in terms of precision. It should be noted that everyone has used their own local datasets and hence no direct comparison is possible. [14, 17, 26, 29] have used broadcast videos for the detection of players. We are performing estimation of number of players rather than just detection. Also we are working on non-commercial videos and not utilizing any temporal information since this is not desirable in ordinary setups where bandwidth and on-site processing power can be problematic. [11] have used thermal cameras for evaluation of their counting algorithm but tested their methodology in indoor sports arenas having closed environment and rela-

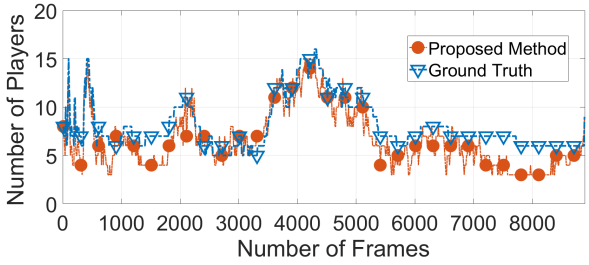



**Fig. 5.10:** Orange line represents the estimated number of people and the blue line shows the ground truth.

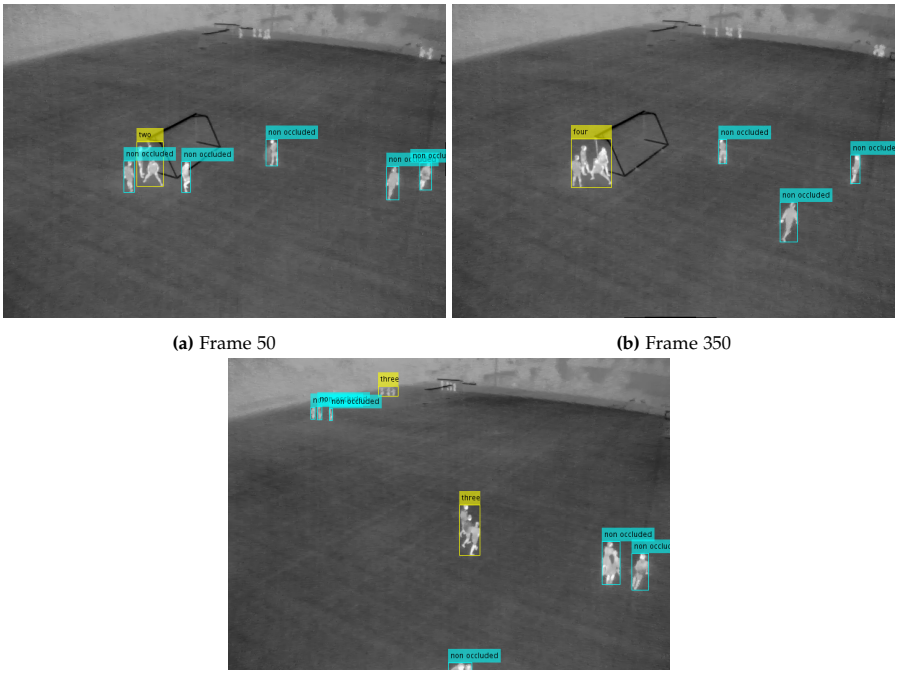
| Method | Acc% | P% | R % |
|---|---|---|---|
| Liu *et al.* [26] | – | 88.6-92.3 | 88.8-92.1 |
| Heydari *et al.* [17] | 96.5 | – | – |
| Manafifard *et al.* [29] | – | 93 | 91 |
| Gade *et al.* [11] | 95.5 | – | – |
| Gerke *et al.* [14] | – | 83-90 | 66-78 |
| **Proposed Method** | **81.4** | **97.8** | **78.8** |

**Table 5.2:** Comparison of our proposed method against previous techniques in terms of Accuracy(Acc), Precision(P) and Recall(R).

tively small area of interest. We are testing our algorithm in a large outdoor soccer field. Better accuracy can probably be achieved by including boundary information and temporal data like in [11].

# 5 Conclusion and Discussion

This paper proposed an automated system for precise counting of players using thermal cameras. A detailed feature vector for each candidate region is formed by using the shape and geometry of the blobs. We used Bagged tree classifier for detection of occlusion. In order to further classify the number of players in occlusion, we proposed a simulation based method. 8990 frames are used for evaluation of the proposed technique for detection of occlusion and estimation of number of players. No ideal conditions are assumed, so it is critical to know that the datasets that we have used contain all types of variations with respect to posture and position of players. The results showed that our proposed method for estimating number of players achieved a high precision, which makes our system suitable for counting precise number of players in groups. Our proposed system is not dependent on




**(a)** Frame 50

**(b)** Frame 350

**(c)** Frame 4000

**Fig. 5.11:** Results after estimation.

light and weather conditions, which make our system more practical for local non-commercial sports analysis.

The mapping from visual appearance of occluding people to the number of individuals could be based on other features than the number of pixels as done in this work. Given sufficient training data more sophisticated handcrafted features or automatically extracted features via a deep learning approach could probably work. But such approaches are likely to require large amounts of annotated data to generalize to arbitrary setups. And since our work is to be applied in many different setups focus has been on a simple feature and an easy training procedure. In fact, for a new setup we need only to input the external camera parameters to the virtual simulation and re-render figure 6.3 and then learn the size of a 1-person blob as a function of distance to the camera in a particular setup. This makes our approach easy to adapt. However, as more fields are analyzed annotated data are automatically collected and future work therefore includes an investigation into the use of deep learning for learning a general mapping from blobs (or bounding boxes) to the number of people [43].

# References

[1] Sermetcan Baysal and Pınar Duygulu. Sentioscope: A soccer player tracking system using model field. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(7):1350–1362, 2016.

[2] M. Beetz, S. Gedikli, J. Bandouch, B. Kirchlechner, N. V. Hoyningen-Huene, and A. C. Perzylo. Visually tracking football games based on tv broadcasts. *20th international joint conference on Artificial intelligence*, 2007.

[3] Leo Breiman. Bagging predictors. *Machine Learning, Springer*, 24(2):123–140, 1996.

[4] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools and Applications, Springer*, 76(3):4405–4425, 2017.

[5] Kyuhyoung Choi and Yongduek Seo. Automatic initialization for 3d soccer player tracking. *Pattern Recognition Letters*, 32(9):1274–1282, 2011.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[7] Cem Direkoglu, Melike Sah, and Noel E. O'Connor. Player detection in field sports. *Machine Vision and Applications, Springer*, 29:187–206, 2018.

[8] Tiziana D'Orazio, Marco Leo, Paolo Spagnolo, Pier Luigi Mazzeo, Nicola Mosca, Massimiliano Nitti, and Arcangelo Distante. An investigation into the feasibility of real-time soccer offside detection from a multiple camera system. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(12):1804–1818, 2009.

[9] Jacob Feldman and Manish Singh. Bayesian estimation of the shape skeleton. *Proceeding of National Academy of Sciences of the USA*, 2006.

[10] Rikke Gade, Anders Jørgensen, and Thomas B. Moeslund. Occupancy analysis of sports arenas using thermal imaging. *International Conference on Computer Vision Theory and Applications*, 2012.

[11] Rikke Gade, Anders Jørgensen, and Thomas B. Moeslund. Long-term occupancy analysis using graph-based optimisation in thermal imagery. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[12] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: a survey. *Machine Vision and Applications, Springer*, 25(1):145–262, 2014.

[13] Rikke Gade and Thomas B. Moeslund. Constrained multi-target tracking for team sports activities. *IPSJ Transactions on Computer Vision and Applications, Springer*, 10(2), 2018.

[14] Sebastian Gerke and Karsten Müller. Identifying soccer players using spatial constellation features. *ACM KDD Workshop on Large-Scale Sports Analytics*, 2015.

[15] Sebastian Gerke, S. Singh, A. Linnemann, and P. Ndjiki-Nya. Unsupervised color classifier training for soccer player detection. *Visual Communications and Image Processing (VCIP)*, 2013.

[16] J. B. Hayet, T. Mathes, J. Czyz, J. Piater, J. Verly, and B. Macq. A modular multi-camera framework for team sports tracking. *Advanced Video and Signal Based Surveillance, IEEE*, 2015.

[17] M. Heydari and A. M. E. Moghadam. An mlp-based player detection and tracking in broadcast soccer video. *International Conference on Robotics and Artificial Intelligence (ICRAI)*, 2012.

[18] Yu Huang, Joan Llach, and Sitaram Bhagavathy. Players and ball detection in soccer videos based on color segmentation and shape analysis. *Multimedia Content Analysis and Mining*, 4577:414–425, 2007.

[19] Naho Inamoto and Hideo Saito. Virtual viewpoint replay for a soccer match by view interpolation from multiple cameras. *IEEE Trans Multimedia*, 9(6):1155–1166, 2007.

[20] Sachiko Iwase and Hideo Saito. Tracking soccer players based on homography among multiple views. *Visual Commun Image Proc*, 5150:283–292, 2003.

[21] J.N.Kapur, P.K.Sahoo, and A.K.C.Wong. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing, Springer*, 29(3):273–285, 1985.

[22] Teddy Ko. A survey on behavior analysis in video surveillance for homeland security applications. *37th IEEE Applied Imagery Pattern Recognition Workshop*, 2008.

[23] Miklas S. Kristoffersen, Jacob V. Dueholm, Rikke Gade, and Thomas B. Moeslund. Pedestrian counting with occlusion handling using stereo thermal camera. *SENSORS*, 62(16), 2016.

[24] Miguel Angel Montañés Laborda, Enrique F. Torres Moreno, Jesús Martínez del Rincón, and José Elías Herrero Jaraba. Real-time gpu color-based segmentation of football players. *Multimedia Tools and Applications*, 73(3):1617–1642, 2012.

[25] G. Liu, D. Zhang, , and H. Li. Research on action recognition of player in broadcast sports video. *International Journal of Multimedia and Ubiquitous Engineering*, 9(10):297–306, 2014.

[26] J. Liu, X. Tong, T. Wang W. Li, Y. Zhang, and H. Wang. Automatic player detection, labeling and tracking in broadcast soccer video. *Pattern Recognition Letters, Elsevier Science Inc*, 30(2):103–113, 2009.

[27] Jingchen Liu, Peter Carr, Robert T. Collins, and Yanxi Liu. Tracking sports players with context-conditioned motion models. In Thomas B. Moeslund, Graham Thomas, and Adrian Hilton, editors, *Computer Vision in Sports*, chapter 6, pages 133–132. Springer, 2014.

[28] SA. Mahmoudi, M. Kierzynka, P. Manneback, and K. Kurowski. Real-time motion tracking using optical flow on multiple gpus. *Bulletin of the Polish academy of sciences, Technical Sciences*, 62(1), 2014.

[29] M. Manafifard, H. Ebadi, and H. Abrishami Moghaddam. Multi-player detection in soccer broadcast videos using a blob-guided particle swarm optimization method. *Multimedia Tools and Applications, Springer*, 76(10):12251–12280, 2016.

[30] Rafael Martín and José M. Martínez. A semi-supervised system for players detection and tracking in multi-camera soccer videos. *Multimedia Tools and Applications*, 73(3):1617–1642, 2014.

[31] I. Kim MM. Khan, TW. Awan and Y. Soh. Tracking occulded objects using kalman filter and color information. *International Journal of Computer Theory and Engineering*, 6(5), 2014.

[32] Thomas B. Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal. *Visual Analysis of Humans: Looking at People*. Springer, Germany, 2011.

[33] Houari Sabirin, Qiang Yao, Keisuke Nonaka, Hiroshi Sankoh, and Sei Naito. Semi-automatic generation of free viewpoint video contents for sport events: Toward real-time delivery of immersive experience. *IEEE MultiMedia*, (99), 2018.

[34] A statistical based analysis of world's most popular sports. Biggest Global Sports. `http://www.biggestglobalsports.com/worlds-biggest-sports/4580873435`. [Online; accessed 03-March-2018].

[35] Satoshi Suzuki. Topological structural analysis of digitized binary images by border. *Computer Vision, Graphics, and Image Processing, Springer*, 30(1):32–46, 1985.

[36] 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, Science Direct*, 159:3–18, 2017.

[37] Pavan Turaga, Rama Chellappa, V. S. Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

[38] Unity Technologies. Unity. `https://unity3d.com/unity`.

[39] Weihong Wang, Jian Zhang, and Chunhua Shen. Improved human detection and classification in thermal images. *IEEE International Conference on Image Processing*, 2010.

[40] Wei Wei and An Yunxiao. Vision-based human motion recognition: A survey. *Second International Conference on Intelligent Networks and Intelligent Systems, IEEE*, 2009.

[41] Ying Yang and Danyang Li. Robust player detection and tracking in broadcast soccer video based on enhanced particle filter. *Journal of Visual Communication and Image Representation Elsevier*, 46:81–94, 2017.

[42] Angela Yao, Dominique Uebersax, and Juergen GallLuc Van Gool. Tracking people in broadcast sports. *Joint Pattern Recognition Symposium, Springer*, 6376:151–161, 2010.

[43] Ho-Sub Yoon, Young lae J. Bae, and Young kyu Yang. A soccer image sequence mosaicking and analysis method using line and advertisement board detection. *ETRI*, 24(6):443–454, 2002.

# Chapter 6

# The Effect of a Diverse Dataset for Transfer Learning in Thermal Person Detection

Noor Ul Huda, Bolette Dybkjær Hansen, Rikke Gade, Thomas B. Moeslund

The paper has been published in the
*Sensors* Vol. 20(7), pp. 1–17, 2020.

*The layout has been revised.*

# Abstract

*Thermal cameras are popular in detection for their precision in surveillance in the dark and for privacy preservation. In the era of data driven problem solving approaches, manually finding and annotating a large amount of data is inefficient in terms of cost and effort. With the introduction of transfer learning, rather than having large datasets, a dataset covering all characteristics and aspects of the target place is more important. In this work, we studied a large thermal dataset recorded for 20 weeks and identified nine phenomena in it. Moreover, we investigated the impact of each phenomenon for model adaptation in transfer learning. Each phenomenon was investigated separately and in combination. the performance was analyzed by computing the F1 score, precision, recall, true negative rate, and false negative rate. Furthermore, to underline our investigation, the trained model with our dataset was further tested on publicly available datasets, and encouraging results were obtained. Finally, our dataset was also made publicly available.*


# 1 Introduction

Person detection is the backbone of many applications ranging from surveillance and military to traffic analysis. Many computer vision branches like behavior analysis, activity recognition, threat recognition, and person re-identification start with the challenge of person detection.

Visual cameras capturing visible light, as well as thermal cameras capturing infrared radiation have been utilized for person detection. Many feature based machine learning [1–4], as well as deep learning [5–7] approaches have been utilized to deal with the problem of person detection in thermal images. Even though thermal cameras have an advantage in outdoor person detection, due to the independence of illumination, robust detection still becomes very challenging in diverse weather and light conditions (see Figure 6.1) and is therefore far from a solved problem.

In the last decade, many deep learning based networks [8–14] have been abundantly created and utilized for person detection in color images. The key to success in the area of machine learning and deep learning is the availability of many datasets [1, 13–16]. Recording and processing of large amount of dataset take much effort and many resources. Alternatively, currently, single shot detectors [8, 10, 11] and transfer learning are also gaining the attention of developers due to their speedy detection and fewer data requirements. Transfer learning refers to learning for a task by transferring the knowledge from the learning of another task. In deep learning, it refers to a method where a model for one task is reused as a starting point for training another task [17]. This reduces the data required, as well as the time needed for training. While learning based approaches have been successful in many com-

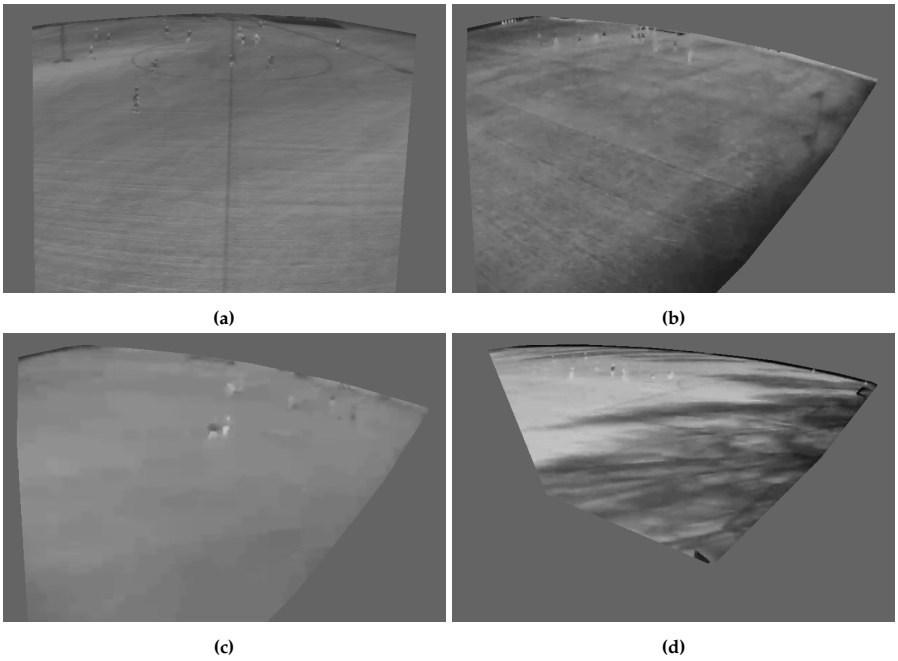



**Fig. 6.1:** Some challenging characteristics in thermal data. (**a**) Varying body temperatures. (**b**) Similar temperatures. (**c**) Motion blur due to wind. (**d**) Shadows.

puter vision and data domains, there is still a large gap in being able to solve thermal detection and classification problems due to the lack of a comprehensive and diverse dataset.

We reviewed the thermal datasets that are available publicly and can be used for person detection. Most of the publicly available thermal datasets (see Table 6.1) are either for tracking or classification. They are short sequences with little variability in the scene, i.e., weather conditions, light conditions, and person heat radiation. This drawback decreases the generalization of detectors. Furthermore, most of the thermal datasets available for person detection are pedestrian data from traffic scenarios and captured from the front view, which makes it difficult to detect people far from the camera. Only one dataset is available that has weather information including haze, rain, and cloudy conditions [18]. However, it contains only a small number of images and hence fails to generalize.

Capturing and annotating a large amount of thermal data are still challenging. An optimal solution would be to study a large range of data and utilize the tool of transfer learning to learn from RGB data. A different range of phenomena affecting thermal videos in the outdoor environment have not been investigated and described yet. Observing the effect of various data

phenomena from thousand of hours of video can help in optimizing dataset development and annotation. The study requires a large dataset recorded over several weeks in different positions and in different places to make sure that all possible outdoor phenomena are covered.

As our first contribution, we studied 20 weeks of variable outdoor thermal data thoroughly to find different phenomena that affect the images. Even by determining all the phenomena, it is still questionable what kind of data are going to have a positive effect and which kind will have a negative effect on person detection in outdoor environments while training a network. Generally, it is presumed that the higher the number of images, the better the detection results. However, due to the high variation of the data characteristics and the low resolution of the thermal images, this is not necessarily the case here, as some phenomena might contribute to a high FP rate. To investigate this research question, as our second contribution, we categorized the phenomena and performed an ablation study for each category. This study gave us a deep analysis of the impact of each category of thermal data and let us choose data in an intelligent manner. This analysis was performed using a single shot deep network and the tool of transfer learning. We employed a single shot deep network due to its high performance and fast learning rate. Finally, the third contribution of this article was a new public thermal dataset for thermal person detection that contains variations regarding the time of day, weather, distance to the camera, various body vs. background temperatures, and shadows. The thermal weights will also be available for researchers for further utilization for transfer learning and solving other thermal data problems.

The rest of the paper is organized as follows: Section 2 provides an overview of the related work. In Section 3, we present our new dataset, and in Section 4, we conduct a thorough investigation into the role of novel training data in transfer learning. Finally, in Section 5, we discuss our findings and future perspectives.

# 2 Related Work

To create an understanding of thermal person detection, the following provides an overview of the state-of-the-art techniques, as well as the datasets used for the evaluation of these techniques.

## 2.1 Multimodal Approaches

Hwang et al. [1] presented a benchmark dataset and baseline code for detection of pedestrians in RGB-Thermal (RGB-T) data. Lahmayed et al. [19] presented a method based on multi-threshold and Histogram of Oriented Gradi-

ents (HOG) and Histograms of Oriented Optical Flow (HOOF) color features combined with an SVM using both thermal infrared and visible light images. They tested their algorithm on the OSU color thermal dataset [20],video analytic dataset [21], and LITIVdataset [22]. Fritz et al. [23] investigated the generalization of a deep learning network in multispectral person detection datasets. They mainly used the Caltech [24], city person [25], CVC-09 [26], KAIST [1], OSU color thermal [20], and Tokyo segmentation [27] datasets for their investigation. Li et al. [28] used the KAISTdataset [1] to create a person detector baseline and then narrowed it down by mining hard negatives. Cuerda et al. [29] employed stream selection based on the confidence map. In this way, they were able to choose the best image out of thermal and visible data based on day and night confidence maps. Many feature extraction and deep learning based approaches have been used for dealing with multimodal data. The problem with multimodal based techniques is the complexity in data handling, as well as the complexity in hardware installation. Here, we are more concerned about thermal only approaches.

## 2.2 Thermal Approaches

Thermal cameras have been utilized in many scenarios ranging from industry to daily life applications [30]. Much research has been carried out for person detection in the infrared domain. Dai et al. [31] presented a method based on background subtraction and shape based classification. They tested their method on the OSU thermal pedestrian database [18]. Zhang et al. [4] also presented a method based on background subtraction and boundary gradients, the temporal coherence of the object area, and the region signature of the intensity distribution. They also tested their method on the OSU thermal database [18]. Li et al. [2] implemented the pedestrian detection in infrared imagery by tuning HOG features. They also tested their algorithm on the OSU thermal pedestrian dataset [18]. A two-stage person recognition approach based on Maximally Stable Extreme Regions (MSERs) and verification of the detected hot spots using a Discrete Cosine Transform (DCT) based descriptor was proposed by Teutsch et al. [3]. They evaluated their approach on the OSU thermal pedestrian [18], OSU color thermal [20], and Terravic motion IR datasets [32]. Many [29, 33–39] used their own datasets for the evaluation.

Recently, Herrmann et al. [5] tested the Single Shot Detector (SSD) with different preprocessing methods to assess thermal performance. They used KAIST [1] for performance evaluation. They [5] also worked with MSERs and CNN and tested on the AMROS, OSU thermal pedestrian [18], OSU color thermal [20], and Terravic motion IR [32] datasets. Tumas et al. [6] proposed an HOG based pedestrian detector combined with CNN for the FIR domain. Heo et al. [7] proposed adaptive Boolean map based saliency combined with YOLO for pedestrian detection at night time. They used CVC-

09 [26] for their experiments. For sports player detection, Gade et al. [37, 38, 40] presented a method based on background subtraction and automatic thresholding. They tested their method on the indoor thermal dataset [40]. Huda et al. [39] previously suggested a simulation based occlusion handling method for detecting and counting the players. This was tested on their own sports dataset.

## 2.3 Datasets

Different multimodal and thermal datasets are publicly available for traffic analysis, surveillance, person tracking, and human pose estimation, among others. The datasets that can be used for person detection are listed in Table 6.1. The scene characteristics, type of data, number of frames, viewpoint, and scene characteristics/or main purpose of the datasets are also provided in the table. All these datasets can be used as pre-training of another network according to the application area.

**Table 6.1:** Available thermal datasets for person detection and the characteristics of each dataset. "Application area/main scene characteristics" summarizes the main features of the videos in each dataset. "Viewpoint" is estimated by generally looking at the image for the camera angle and the distance of persons from the camera.

| **Name** | **# of Frames** | **# of Seq** | **Viewpoint** | **Application Area/ Main Scene Characteristics** | **Camera/ Image Specifications** |
|---|---|---|---|---|---|
| KAIST [1] | 95 k | | Near front | outdoor traffic, day and night multispectral | 640 × 480, 20 Hz |
| OSU Color Thermal (CT) [20] | 17 k | | Far top | Outdoor walkway | Raytheon PalmIR 250D, 320 × 240, 30 Hz |
| AAU-VAP TPD [41] | 5.7 k | 3 | Near front | Indoor office | Axis Q1922 640 × 480 30 Hz |
| LITIV- -VAP [22] | 4.3 k | | Near front | Indoor hall | |
| CVC-09 [26] | 11 k | | Near | Traffic pedestrian, day and night | 640 × 480 |
| CVC-14 [42] | 7.7 k | | Near | Traffic pedestrian, day and night | |

**Table 6.1:** *Cont.*

| Name | # of Frames | # of Seq | Viewpoint | Application Area/ Main Scene Characteristics | Camera/ Image Specifications |
|---|---|---|---|---|---|
| LITIV- -2018 [43] | | 3 | Near front | Indoor hall | |
| OSU Thermal (T) [18] | 0.2 k | | Far top | Outdoor pedestrian haze, fair, light rain, partially cloudy | Raytheon 300D, 320 × 240, 30 Hz |
| ASL-TID [44] | 4.3 k | 8 | Varied | Outdoor varied background, person, cat, horse | FLIR Tau 324 × 256 |
| Terravic Motion IR [32] | 23.7 k | 18 | Varied | Outdoor tracking, surveillance, indoor hallway, plane tracking, underwater and near-surface motion, background motion | Raytheon L-3 Thermal-eye 2000AS, 320 × 240 |
| LSI Dataset [45] | 15.2 k | 13 | | Outdoor pedestrian Hz | Intigo Omega imager, 164 × 129 |
| BU-TIV [46] Benchmark | 7.9 k | 2 | Near | Indoor atrium | 512 × 512 |
| Atrium | | | | | |
| Lab | 26.7 k | 3 | Near | Indoor and outdoor marathon | 512 × 512 |
| Marathon | 1 k | | Very far | | 1024 × 640 |
| VOT-TIR 2015 [49] | 270 | 1 | Near front | Fair outdoor | 640 × 480, 30 Hz |
| Birds | | | | | |
| Crossing | 301 | 1 | Near top | Fair outdoor | 640 × 480, 30 Hz |
| Crouching | 618 | 1 | Near front | Outdoor roadside | 640 × 480, 30 Hz |
| Crowd | 71 | 1 | Near front | Outdoor roadside occluded | 640 × 512, 30 Hz |
| Street | 172 | 1 | Far front | Outdoor street | 640 × 480, 30 Hz |

**Table 6.1:** *Cont.*

| Name | # of Frames | # of Seq | Viewpoint | Application Area/ Main Scene Characteristics | Camera/ Image Specifications |
|---|---|---|---|---|---|
| Saturated | 218 | 1 | Near front | Outdoor street occluded | 640 × 480, 30 Hz |
| Mixed distractor | 270 | 1 | Near front | Indoor | 527 × 422, 30 Hz |
| Hiding | 358 | 1 | Near front | Indoor | 263 × 210, 30 Hz |
| Garden | 676 | 1 | Near top | Outdoor garden | 324 × 256, 30 Hz |
| Depth-wise crossing | 851 | 1 | Medium top | Outdoor fair roadside | 640 × 480, 30 Hz |
| Trees | 665 | 1 | Far top | Outdoor dark | 640 × 480, 30 Hz |
| Thermal soccer dataset [40] | 3000 | 4 | Near top | Indoor soccer arena | 640 × 480, 30 Hz |

All the datasets available consisted of sequences with a short duration; thus, they had less variability in terms of weather and light conditions. Most of the available datasets were pedestrian data from traffic data analysis and captured from a frontal viewpoint. Many datasets were indoor, and thus, these were captured in controlled light and temperature conditions and did not include all the variability of outdoor environments. Even with a large number of frames [1] and weather information [18], it was still questionable if the data were enough to include all outdoor phenomena. Therefore, the research community lacks a comprehensive and diverse dataset to develop robust algorithms for the detection of people. Therefore, we studied long durations of data and came up with a shorter, but novel and diverse dataset below that is comprised of all outdoor phenomena.

# 3 Novel Dataset

The first contribution of this paper is the investigation and study of a diverse thermal dataset for person detection. In thermal images, weather conditions have a similar effect as lighting conditions have on RGB images. it is therefore essential to include varying weather and light effects in a dataset. Furthermore, because the resolution of thermal sensors is still relatively low, the size of objects in images is also an important factor. The data we recorded were captured in outdoor sports fields with people playing soccer or performing

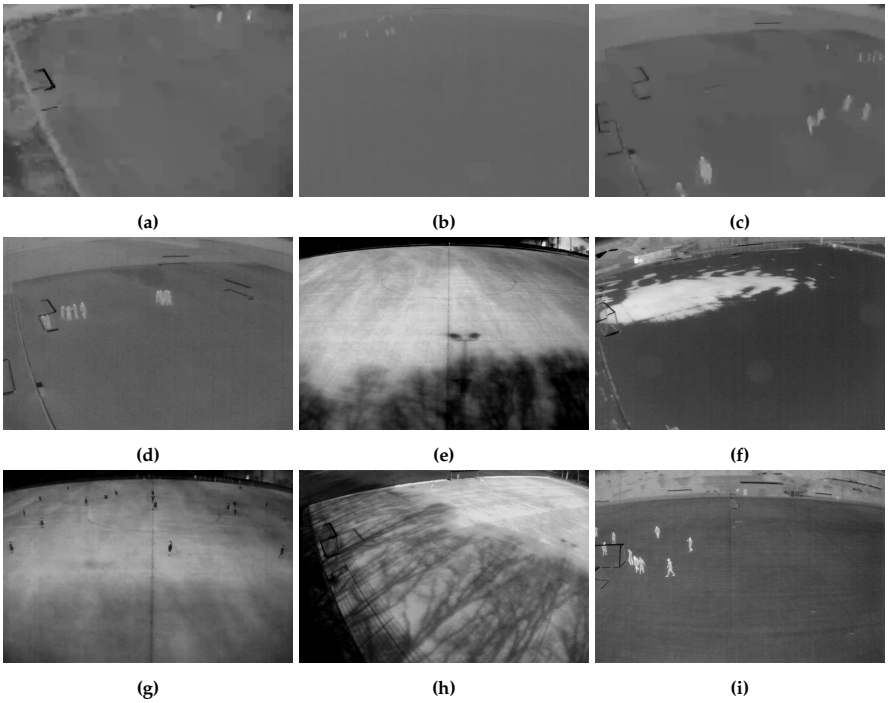



**Fig. 6.2:** Nine different phenomena that are included in our novel dataset. (**a**) Low resolution. (**b**) Far viewpoint. (**c**) Wind. (**d**). Occlusion. (**e**) Shadow. (**f**) Snow. (**g**) Opposite temperature. (**h**) Similar temperature. (**i**) Good condition.

related exercises. The nature of these recordings ensured that many challenges related to person detection were included: different scales, pose variations, interactions/occlusions between people, and fast and erratic motion. Regarding the weather effects, we recorded 20 weeks of thermal recordings across January to April in Denmark. Therefore, it spanned the periods from little daylight to bright sunny days and snowy days of winter to pleasant spring days. In the recordings, we experienced several different key challenges: varying temperatures (people hotter/colder/same temperature than the ground), shadows (parts of the ground were not heated by the Sun), wind (camera moving), snow (regions on the ground with different reflection and emissivity of heat), and occlusion (people in groups) in the thermal images.

After examining all challenges and scrutinizing the entirety of the data, we suggested that nine different phenomena should be included in a dataset for it to be sufficiently diverse and help the model generalize outdoor person detection in thermal images. These nine phenomena are listed and illustrated in Figure 6.2.

### 3.1 Data Recording

We recorded thermal videos from 10 different sports fields for two weeks each, which comprised 20 weeks of data. The cameras used for recording were Axis Q1921 (resolution 384 $\times$ 288 pixels) and Axis 1922 (resolution 640 $\times$ 480), and they were mounted approximately 9m above the ground on a light pole surrounding the field. Three cameras were installed at the center of each field to cover the entire field area. The sequences selected for this investigation were from all of the cameras' views. The recordings were done from January 2018 to April 2018.

### 3.2 Data Description

As the first step in transfer learning is a model adaptation, we used 3000 indoor publicly available images [40] as pre-training images for model adaptation. The dataset from [40] was selected for pre-training as it had nearly perfect thermal data, i.e., lighter person on a darker background. Moreover, it was similar to our dataset as it was recorded in an indoor sports field and contained 24,000 person annotations. As the data from [40] helped in model adaptation and saved in annotation cost, our new dataset (Table 6.2) helped in obtaining the goal of generalization in detection as it included all possible outdoor phenomena from an outdoor environment.

Manually annotating all the data was unrealistic. Therefore, we scrutinized the periods where all nine phenomena occurred, and the number of players in a given image in these periods varied (from 0 to 40). In each period, we selected a frame every 160th second and annotated that frame. This large temporal gap between annotated frames was introduced to enforce as much diversity as possible. One-thousand nine-hundred forty-one frames were selected as the training dataset. In these frames, a total of 5590 persons were annotated. The details of the dataset are presented in Table 6.2. For testing purposes, 1000 more frames were randomly selected from all the recorded data (100 frames from two weeks of video). it was manually checked that no image from the training data was repeated in the testing data. The camera view (left, right, middle) was also selected randomly. All of the data were annotated with the MATLAB object detection bounding box annotator [47]. Our person detection dataset (PD-T) is available at `http://www.vap.aau.dk/dataset/`.

## 4 Investigating the Role of Training Data

A traditional deep learning network contains a large number of parameters. Training such a network requires an enormous amount of training

**Table 6.2:** Key characteristics of the proposed training data.

| Category | Phenomena | # of Frames | # of Persons |
|---|---|---|---|
| Viewpoint | Good condition | 122 | 632 |
| | Far viewpoint | 64 | 652 |
| Heat effects | Opposite temperature | 72 | 792 |
| | Similar temperature | 107 | 644 |
| Image artifacts | Low resolution | 158 | 734 |
| | fOcclusion | 20 | 305 |
| Weather effects | Shadow | 171 | 742 |
| | Snow | 1060 | 168 |
| | Wind | 167 | 921 |

data. The online availability of such an enormous amount of data is not always a possibility, especially in non-RGB applications. Transfer learning is the optimal solution in such conditions since many features in the first layers of a deep learning network are similar across applications [48]. The question is which phenomena need to be included in a dataset for outdoor thermal person detection for a positive transfer. To investigate this research question, we needed a pre-trained detection algorithm on which we could apply transfer learning with our data. we chose the CNN based single shot detector YOLOv3 [8].

You Only Look Once (YOLO) is one of the fastest deep learning algorithms for the detection of objects in an image, which can process 45 frames per second. This algorithm treats the problem of detection as a regression problem and trains on the whole image at once to optimize the performance. Moreover, it detects the class objects with their probabilities at the same time without requiring region proposals.

The YOLOv3 network, used in this work, divided every training image into a grid of ($S \times S$) cells. it searched for the center of the target objects in these grid cells. $B$ number of bounding boxes with their confidence scores could be predicted by each grid cell. Confidence was defined as the probability of detected objects multiplied by the Intersection over Union (IoU) between the ground truth bounding box area and the detected object bounding box area.

The model was more effective at detecting small objects compared to previous versions of YOLO because it predicted bounding boxes at different scales. This added multiscale detection in v3 allowed us to detect a person very far from the camera. At the same time, the number of predictable bounding boxes in each cell provided some limitation on the detection.

**Table 6.3:** List of combinations for tests.

| # | Combinations | # | Combinations |
|---|---|---|---|
| 1. | Indoor | 9. | Indoor+heat effects +image artifacts |
| 2. | Indoor+viewpoint | 10. | Indoor+heat effects +weather effects |
| 3. | Indoor+heat effects | 11. | Indoor+image artifacts +weather effects |
| 4. | Indoor+image artifacts | 12. | Indoor+Viewpoint +heat effects+image artifacts |
| 5. | Indoor+weather effects | 13. | Indoor+viewpoint +heat effects+weather effects |
| 6. | Indoor+viewpoint +heat effects | 14. | Indoor+heat effects +image artifacts+weather effects |
| 7. | Indoor+viewpoint +image artifacts | 15. | Indoor+viewpoint +image artifacts+weather effects |
| 8. | Indoor+viewpoint +weather effects | 16. | Indoor+viewpoint+heat effects +image artifacts+weather effects |

## 4.1 Assessment Protocol

To assess the role of training data, we divided our training data based on the phenomena discussed in Section 3 into categories defined in Table 6.2. The amount of test data was always kept the same. Tests were performed by adding one category of images at a time and then combining different categories of images. A total of 16 different combinations were tested, listed in Table 6.3. Indoor data were from [40] and were used as a baseline for model adaptation. Results for each of these combinations would provide insights into how different types of training data affected the detection results on varying data.

For transfer learning, we used convolution weights that were pre-trained on ImageNet [14] using the Darknet53 [8] model due to their reported high performance and speed [8]. The network was trained with S = 7, where network iterations were set to 40,000, and the results from the mean of iterations (10,000, 20,000, 30,000, and 40,000) were considered. Here, we set the learning rate to 0.001, momentum to 0.9, and decay to 0.0005. The training and testing of all combinations were performed using a graphical processing unit GTX 1080 with Linux Ubuntu 16.04.

## 4.2 Evaluation

We used precision, recall, F1 score, False Negative Rate (FNR), and True Negative Rate (TNR) as the performance measures. Along with recall and precision, we were also interested in true and false negative rates, as these matrices are of great importance in surveillance and occupancy analysis applications, where an event of negative detection is as important as an event as positive detection. The F1 scores of all the combinations are provided in Table 6.4. Recall, precision, TNR, and FNR are illustrated in Figure 6.3. Here, we calculated our measures, i.e., F1 score, recall, precision, TNR, and FNR, as:

$$\text{F1 score} = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6.1}$$

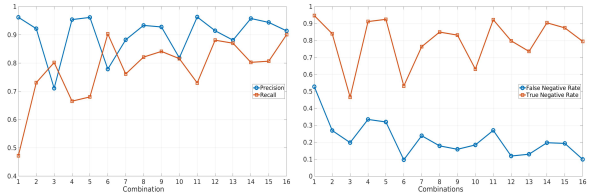
$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FN}}, \quad \text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{6.2}$$

True Positives (TP) were defined as the number of persons that were correctly detected as persons and True Negatives (TN) as the number of images with zero persons correctly identified as having zero persons. False Positives (FP) represented the regions in the image with no person, but there was nonetheless a person detected. False Negatives (FN) represented the regions where persons were present, but the detector failed to recognize them.

Results presented in Table 6.4 indicated that for Combinations 2 to 5, when only one category was added at a time, viewpoint images significantly increased the value of the F1 score, indicated by green, while the images with the heat effect had the least impact on the results, indicated by red. For Combinations 6 to 11, the alliance of heat and weather effects and the alliance of viewpoint and image artifacts seemed to have the lowest performance. The combinations of heat effect and image artifacts and the combination of viewpoint and weather effects had the highest performance in terms of F1 score. For the last combinations, 12–15, we could see that including all categories exclusive of the weather effect had the highest F1 score of 89.74%, while the other combinations performed almost equally. The last combination with all data included as expected showed the maximum performance in terms of F1 score.

In looking individually at the results of each combinations, one noticeable observation was found with Combinations 2, 7 and 10. These combinations almost had the same performance. Although, if we looked at the number of images in Combinations 2 and 10, Combination 10 had more than three times the number of images as Combination 2. The same pattern could be observed in Combinations 12 and 16. The weather effect contained more than half of the data, but its inclusion increased the performance only by 1%.

The overall contribution of each category is also shown in the last row of Table 6.4. The mean was computed by taking the mean of all F1 scores




**(a)** **(b)**

**Fig. 6.3:** (**a**) Precision and Recall, (**b**) True Negative Rate (TNR) and False Negative Rate (FNR).

in which a particular category was included. Results were consistent with the precision and TNR results, and heat effects had the lowest F1 score. The highest F1 score was obtained for the viewpoint category, which had images with good contrast and both far and close views. Moreover, this category introduced scene adaptation from an indoor to outdoor field environment. it could also be observed that although the image artifacts category had eight times fewer training images than weather effects, it had a better mean F1 score.

The results obtained from the experiment are also presented in Figure 6.3. Precision and recall are shown in Figure 6.3a, and FNR and TNR are shown in Figure 6.3b. it can be seen that for certain combinations, i.e., 3, 6, 10, and 13, there were visible dips in the precision and TNR values. The magnitude of the dip in precision was less than the TNR because only FP was considered in the calculation of precision, whereas in the TNR calculation, both FP and TN played a role.

If we looked at all these combinations, the common category was "heat effects". The other noticeable effect was the decrease in the dip magnitude with the addition of more categories. As more and more categories were added to "heat effects", the precision and TNR both improved. There was no significant change observed in the FNR results. However, the recall had an opposite effect from the precision and TNR, as the addition of the "heat effects" category improved recall. The details of this improvement are explained later in the section.

The precision and TNR were maximum for the image artifacts and weather effects categories. This was because occlusion and low resolution images were present in the image artifacts category, and the FP and FN reduced; whereas for weather effects, more images of empty fields with snow and shadow were added in the training data. Snow and shadow could sometimes resemble humans and be detected as persons. Therefore, with the addition of the weather effect category, FNR and TNR both improved.

Herrmann et al. concluded that an inverted thermal dataset had a resemblance to the grayscale of RGB data. Therefore, the domain adaptation was quicker when pretrained RGB weights were used. In our results, we could

**Table 6.4:** F1 score of combinations. Here, **X** indicates the category added in a combination. Other than indoor data, Combinations 2–5 only had one category of images, Combinations 6–11 two categories of images, and Combinations 12–15 three categories of images. Lastly, Combination 16 had all categories. The red color shows the worst-performing combinations, and the green color shows the best performing combinations within each section.

| Combinations | Indoor | Viewpoint | Heat Effects | Image Artifacts | Weather Effects | F1 Score |
|---|---|---|---|---|---|---|
| **1** | X | | | | | 63.35 |
| **2** | X | X | | | | 81.52 |
| **3** | X | | X | | | 75.35 |
| **4** | X | | | X | | 78.39 |
| **5** | X | | | | X | 79.68 |
| **6** | X | X | X | | | 83.63 |
| **7** | X | X | | X | | 81.74 |
| **8** | X | X | | | X | 87.37 |
| **9** | X | | X | X | | 88.24 |
| **10** | X | | X | | X | 81.71 |
| **11** | X | | | X | X | 83.04 |
| **12** | X | X | X | X | | **89.74** |
| **13** | X | X | X | | X | 87.57 |
| **14** | X | | X | X | X | 87.34 |
| **15** | X | X | | X | X | 86.99 |
| **16** | X | X | X | X | X | **90.23** |
| Mean | 82.87 | 86.10 | 85.48 | 85.78 | 85.49 | |

also observe a similar response in terms of recall.

We can see in Figure 6.3a that every time the heat effects category was added, recall improved. However, at the same time, precision and TNR reduced. All the other categories in Table 6.2, except heat effects, had images with persons in the dark background. Therefore, the heat effect category, which was 8% of the complete training dataset, acted as noise. In particular, similar temperature images had the most effect on reducing TNR. Any lesser contrast noise could be detected as FP. This problem could be solved by generalizing the dataset in a single domain by detecting the heat category events. Results also suggested that converting the whole dataset into inverted thermal images might be more beneficial, as this would help improve the recall and model adaptation.

To select which category to include in training, it still depended on the target application. For example, if we compared Combinations 12 and 16, the increase in the F1 score was only 0.49% by including the data from the weather effect category. To show the effect of including the weather effect data, a few test images are shown in Figure 6.4. Figure 6.4a,b is from our dataset, and Figure 6.4c–d were taken from the publicly available CVC-09 database. Figure 6.4a,c,e was tested with Combination 16, where the weather effect was included; whereas Figure 6.4b,d,f is the results of the same images when the weather effect was not included, i.e., Combination 12. it can be seen that without the weather effect, TN and FN were better; however, with its inclusion, TP improved, but the FPR also increased. For example, if we needed the system for surveillance, then it would be important to avoid an FN event. In such cases, weather effects data would be required for training. Occupancy analysis has similar requirements.

## 4.3 Results on Publicly Available Datasets

We picked three public datasets to test the generalization of our trained weights for person detection. The datasets consisted of three different diverse datasets from Table 6.1: CVC-09 [26], OSU-T [18], and BU-TIV-atrium [46].

OSU-T was recorded outdoors with different weather conditions, as mentioned in Table 6.1. it consisted of 284 images. The data were captured from a far top viewpoint. CVC-09 was recorded from a camera in a car while driving. The images were divided into two subsets for day and night. CVC-09 (day) consisted of 2881 test images and 4223 training images, out of which 1112 were negative frames and 3111 positive frames. CVC-09 (night) consisted of 2883 test images and 3200 training images, out of which 1001 were negative frames and 2199 positive frames. BU-TIV was recorded indoors with a near top viewpoint. it had three sequences of videos with Views 1, 2, and 3. we chose its View 1 for our tests, which consisted of 3482 images.

Tests on publicly available datasets were performed in two sessions. Firstly

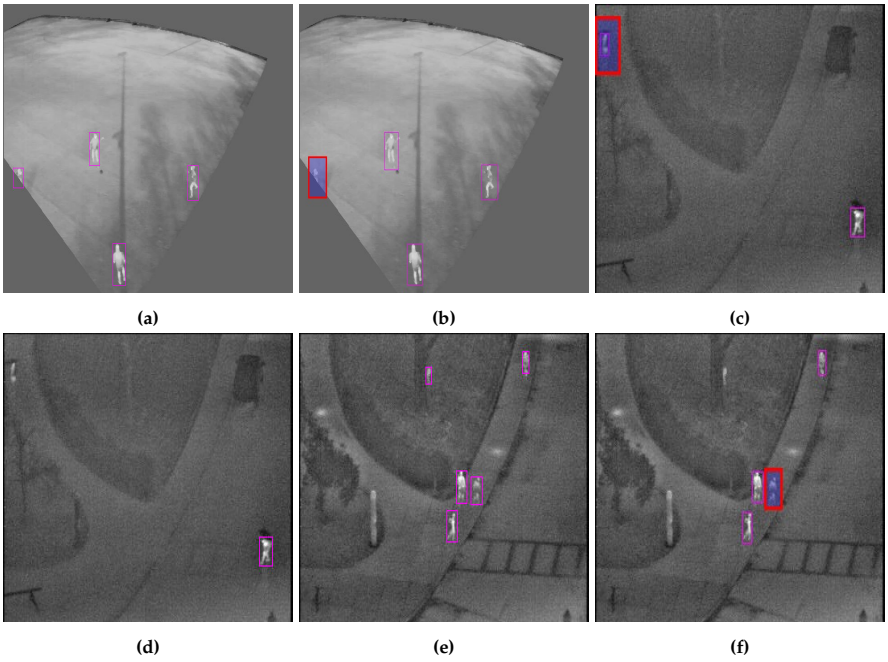
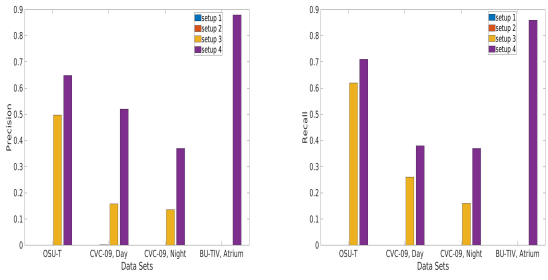



**Fig. 6.4:** Example images for qualitative assessment. (**a,c,e**) are the results for Combination 16, while (**b,d,f**) are the results for the same images from Combination 12. The images (**a,b**) are from our test data, and the images (**c–f**) are from OSU-T [18]. In these images, highlighted red boxes are incorrect detections.




**(a)** **(b)**

**Fig. 6.5:** (**a**) Precision and (**b**) recall measures of different training weights on publicly available datasets. Here, the blue bars are the results tested by our thermal training weights, and orange bars are the results tested by our thermal training weights and further training by adding only 5% of the new dataset for 100 iterations.

the images were tested using the weights obtained from Combination 16, shown in Table 6.3. In the second session, tests were performed by adding 5% of the data from the public dataset to the Combination 16 dataset and retraining it.

For training the second session test, from OSU-T and BU-TIV, we added 5% of the whole data in training corresponding to 14 and 174 images, respectively, and from CVC-09 (day and night), 5% of the training data was added to the training set corresponding to 211 and 160 images, respectively. The number of iterations for learning was 100 to avoid overfitting due to a small number of training images.

Results of this experiment are presented in Figure 6.5. Blue bars are the results obtained from Combination 16 weights, and red bars are the results obtained after retraining Combination 16 with 5% of the public dataset. it can be seen that by using the weights from Combination 16, the performance was not good, and in the case of BU-TIV, the algorithm failed to detect anything. In BU-TIV, the viewpoint was different, and people appeared larger than in our dataset. However, with only 5% of training data and with 100 iterations, a significant increase in precision could be seen. The highest precision was obtained for BU-TIV and the lowest for the CVC data, with an average precision of 0.69%. In BU-TIV and OSU-T, there were no other heated objects present other than humans, and in OSU-T, the viewpoint was very similar to our dataset; therefore, good precision results were achieved.

In the CVC dataset, a significant difference between day and night results was observed. During the day, the temperatures of car bodies, tires, and other objects increased. Their pattern became similar to human body features, which increased FPR and decreased precision. Example results from all datasets used for evaluations are shown in Figure 6.6.

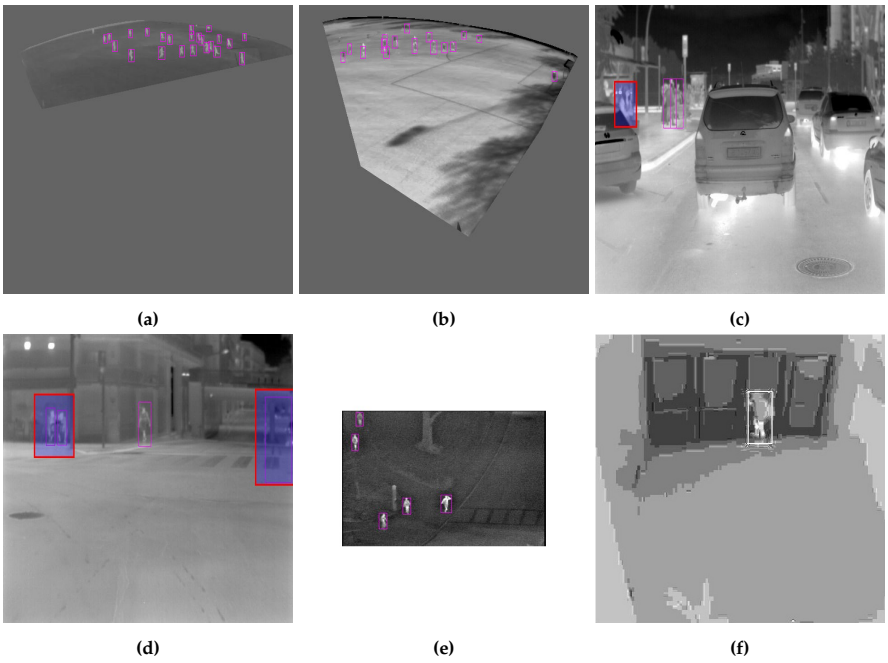



**Fig. 6.6:** Example images for qualitative assessment. The images (**a**) and (**b**) are from our test data. The results of images (**a**) and (**b**) are obtained from Combination 16, shown in Table 6.4. Image (**c**) is from CVC-day [26], image (**d**) from CVC-night [26], image (**e**) from OSU-T [18], and image (**f**) from BU-TIV-atrium [46]. The contrast of (**f**) is adjusted for better visualization. In these images, highlighted red boxes are wrong detections.

# 5 Conclusions

In this work, we reviewed publicly available thermal datasets that could be used for person detection, and we documented the lack of diversity in these datasets. we also studied and presented a new thermal dataset and found nine different phenomena that could occur in outdoor soccer fields. The phenomena were further categorized into four categories. The impact of each category was studied for model generalization using transfer learning. Results showed that each category benefited the model generalization differently. The results showed that depending on the application, categories could be selected intelligently to obtain the desired results. The weights obtained from our dataset were further tested on three publicly available datasets. For a relatively small amount of training data from a new domain and with few iterations, good performance was achieved for person detection. Results showed that our weights could be used for model adaptation for a new domain. This will help researchers save the effort of annotating large datasets and also the time for training a new network from scratch. Moreover, with weights for YOLOv3, our new dataset is made publicly available for further research.

# References

[1] Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.

[2] Li, W.; Zheng, D.; Zhao, T.; Yang, M. An effective approach to pedestrian detection in thermal imagery. In Proceedings of the International Conference on Natural Computation, Chongqing, China, 29–31 May 2012; pp. 325–329, doi:10.1109/ICNC.2012.6234621.

[3] Teutsch, M.; Mueller, T.; Huber, M.; Beyerer, J. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Columbus, OH, USA, 24–27 June 2014; pp. 209–216, doi:10.1109/CVPRW.2014.40.

[4] Zhang, H.; Zhao, B.; Tang, L.; Li, J. Variational based contour tracking in infrared imagery. In Proceedings of the International Congress on Image and Signal Processing, Tianjin, China, 17–19 October 2009; pp. 1–5, doi:10.1109/CISP.2009.5303802.

[5] Herrmann, C.; Müller, T.; Willersinn, D.; Beyerer, J. *Real-Time Person Detection in Low-Resolution Thermal Infrared Imagery with MSER and CNNs*; SPIE: Bellingham, WA, USA, 201; Volume 9987, doi:10.1117/12.2240940.

[6] Tumas, P.; Jonkus, A.; Serackis, A. Acceleration of HOG based pedestrian detection in FIR camera video stream. In Proceedings of the IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 26 April 2018; pp. 1–4, doi:10.1109/eStream.2018.8394126.

[7] Heo, D.; Lee, E.; Ko, B.C. Pedestrian detection at night using deep neural networks and saliency maps. *Electron. Imaging* **2018**, *2018*, 060403–1, doi:doi.org/10.2352/J.ImagingSci.Technol.2017.61.6.060403.

[8] Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.

[9] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.

[10] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin, Germany, 2016; pp. 21–37.

[11] Cioppa, A.; Deliège, A.; Van Droogenbroeck, M. A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1765–1774, doi:10.1109/CVPRW.2018.00229.

[12] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 580–587.

[13] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.

[14] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Lake Tahoe, NV, USA, 2012; pp. 1097–1105.

[15] Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. the pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338, doi:10.1007/s11263-009-0275-4.

[16] Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September, 2014; Springer: Berlin, Germany, 2014; pp. 740–755.

[17] Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, doi:10.1186/s40537-016-0043-6.

[18] Davis, J.W.; Keck, M.A. A two-stage template approach to person detection in thermal imagery. In Proceedings of the IEEE Workshops on Applications of Computer Vision, Breckenridge, CO, USA, 5–7 January 2005, Vol. 1, pp. 364–369, doi:10.1109/ACVMOT.2005.14.

[19] Lahmyed, R.; El Ansari, M.; Ellahyani, A. A new thermal infrared and visible spectrum images based pedestrian detection system. *Multimed. Tools Appl.* **2019**, *78*, 15861–15885, doi:10.1007/s11042-018-6974-5.

[20] Davis, J.W.; Sharma, V. Background-subtraction using contour based fusion of thermal and visible imagery. *Comput. Vis. Image Underst.* **2007**, *106*, 162–182, doi:10.1016/j.cviu.2006.06.010.

[21] Video Analytics Dataset. Available online: `https://www.ino.ca/en/technologies/video-analytics-dataset/`. (accessed on 26 June 2019).

[22] Torabi, A.; Massé, G.; Bilodeau, G.A. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Comput. Vis. Image Underst.* **2012**, *116*, 210–221, doi:10.1016/j.cviu.2011.10.006.

[23] Fritz, K.; König, D.; Klauck, U.; Teutsch, M. *Generalization Ability of Region Proposal Networks for Multispectral Person Detection*; SPIE Defense + Commercial Sensing, Baltimore, MD, USA, 2019; Volume 10988, doi:10.1117/12.2520705.

[24] Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761, doi:10.1109/TPAMI.2011.155.

[25] Zhang, S.; Benenson, R.; Schiele, B. CityPersons: a diverse dataset for pedestrian detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4457–4465, doi:10.1109/CVPR.2017.474.

[26] Socarras, Y.; Ramos, S.; Vázquez, D.; López, A.; Gevers, T. Adapting pedestrian detection from synthetic to far infrared images. In Proceedings of the International Conference on Computer Vision (ICCV) Workshop, Sydney, Australia, 1–8 December 2013.

[27] Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Vancouver, BC, Canada, 24–28 September 2017; pp. 5108–5115, doi:10.1109/IROS.2017.8206396.

[28] Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral pedestrian detection via simultaneous detection and segmentation. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle , UK, 3–6 September 2018.

[29] Cuerda, J.S.; Caballero, A.F.; López, M.T. Selection of a visible-light vs. thermal infrared sensor in dynamic environments based on confidence measures. *Appl. Sci.* **2014**, pp. 331–350, doi:10.3390/app4030331.

[30] Gade, R.; Moeslund, T.B. Thermal cameras and applications: a survey. *Mach. Vis. Appl.* **2013**, *25*, 245–262.

[31] Dai, C.; Zheng, Y.; Li, X. Layered representation for pedestrian detection and tracking in infrared imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; doi:10.1109/CVPR.2005.483.

[32] Miezianko, R.; Pokrajac, D. People detection in low resolution infrared videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Anchorage, AK, USA, 23–28 June 2008; pp. 1–6, doi:10.1109/CVPRW.2008.4563056.

[33] Jungling, K.; Arens, M. Feature based person detection beyond the visible spectrum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Miami, FL, USA, 20–25 June 2009; pp. 30–37, doi:10.1109/CVPRW.2009.5204085.

[34] Teutsch, M.; Müller, T. Hot spot detection and classification in LWIR videos for person recognition. In *Automatic Target Recognition XXIII*; SPIE: Bellingham, MA, USA, 2013; Volume 8744, doi:10.1117/12.2015754.

[35] Wang, J.; Bebis, G.; Miller, R. Robust video based surveillance by integrating target detection with tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop, New York, NY, USA, 17–22 June 2006; pp. 137–137, doi:10.1109/CVPRW.2006.180.

[36] Zhang, L.; Wu, B.; Nevatia, R. Pedestrian detection in infrared images based on local shape features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8, doi:10.1109/CVPR.2007.383452.

[37] Gade, R.; Jørgensen, A.; Moeslund, T.B. Long-term occupancy analysis using Graph-Based Optimisation in thermal imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June IEEE Computer Society Press: Los Alamitos, CA, USA, 2013; pp. 3698–3705, doi:10.1109/CVPR.2013.474.

[38] Gade, R.; Jørgensen, A.; Moeslund, T.B. Occupancy analysis of sports arenas using thermal imaging. In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), Rome, Italy, 24–26 February 2012; SCITEPRESS Digital Library: Setubal, Portugal, 2012; pp. 277–283.

[39] Huda, N.; Jensen, K.; Gade, R.; Moeslund, T. Estimating the Number of Soccer Players using Simulation based Occlusion Handling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1937–1946, doi:10.1109/CVPRW.2018.00236.

[40] Gade, R.; Moeslund, T.B. Constrained multi-target tracking for team sports activities. *IPSJ Trans. Comput. Vis. Appl.* **2018**, doi:10.1186/s41074-017-0038-z.

[41] Palmero, C.; Clapés, A.; Bahnsen, C.; Møgelmose, A.; Moeslund, T.B.; Escalera, S. Multi-modal rgb–depth–thermal human body segmentation. *Int. J. Comput. Vis.* **2016**, *118*, 217–239.

[42] González, A.; Fang, Z.; Socarras, Y.; Serrat, J.; Vázquez, D.; Xu, J.; López, A.M. Pedestrian detection at day/night time with visible and FIR cameras: a comparison. *Sensors* **2016**, *16*, 820, doi:10.3390/s16060820.

[43] St-Charles, P.L.; Bilodeau, G.A.; Bergevin, R. Online mutual foreground segmentation for multispectral stereo videos. *Int. J. Comput. Vis.* **2019**, *127*, 1044–1062.

[44] Portmann, J.; Lynen, S.; Chli, M.; Siegwart, R. People detection and tracking from aerial thermal views. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 1794–1800, doi:10.1109/ICRA.2014.6907094.

[45] Olmeda, D.; Premebida, C.; Nunes, U.; Armingol, J.M.; Escalera, A.d.l. *LSI far Infrared Pedestrian Dataset*; Universidad Carlos III de Madrid: Madrid, Spain, 2019, doi:10.21950/VBIIBU.

[46] Wu, Z.; Fuller, N.W.; Theriault, D.H.; Betke, M. A thermal infrared video benchmark for visual analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 201–208.

[47] Image labeler MATLAB 2019. Available oneline: `http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htmhttps://www.mathworks.com/help/vision/ug/get-started-with-the-image-labeler.html` (accessed on 3 November 2019).

[48] Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *arXiv* **2014**, arXiv:1411.1792.

[49] Felsberg, M.; Berg, A.; Hager, G.; Ahlberg, J.; Kristan, M. ; others, The thermal infrared visual object tracking VOT-TIR2015 challenge results. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, CH, USA, 7–13 December 2015; pp. 76–88.

# Chapter 7

# Effects of Pre-processing on the Performance of Transfer Learning Based Person Detection in Thermal Images

Noor Ul Huda, Rikke Gade, Thomas B. Moeslund

The paper has been published in the
International Conference on Pattern Recognition and Machine Learning
(PRML).

*The layout has been revised.*

# Abstract

*Thermal images have the property of identifying objects even in low light conditions. However, person detection in thermal is tricky, due to varying person representations depending upon the surrounding temperature. Three major polarities are commonly observed in these representations i.e., 1. person warmer than the background, 2. person colder than the background and 3. person's body temperature is similar to background. In this work, we have studied and analyzed the performance of the detection network by using the data in its original form and by harmonizing the person representation in two ways i.e., dark persons in the light background and light persons in a darker background. The data passed to each testing scenario was first pre-processed using histogram stretching to enhance the contrast. The work also presents the method to separate the three kinds of images from thermal data. The analysis is performed on publicly available outdoor AAU-PD-T and OSU-T datasets. Precision, recall, and F1 score is used to evaluate network performance. The results have shown that network performance is not enhanced by performing the mentioned pre-processing. Best results are obtained by using the data in its original form.*


# 1 Introduction

Person detection is the fundamental problem in any human-related computer vision based approach. Different camera setups, including RGB and thermal, have been proposed for person detection.
In RGB domain, well-defined solutions based on machine learning and deep learning have been reported. However, RGB cameras have their limitation in low light and total blackout conditions. Whereas thermal cameras have the advantage to perform very well in such situations. On the contrary, the outcome from the thermal cameras gets affected by the temperature of the surrounding objects as well as the temperature of the environment.
Most of the reported solutions in thermal domain are based on finest videos and image-based data experiments. The data used had the characteristics of higher person temperature vs lower background temperature, along with small occlusion, less reflection and weather effects. Therefore, It becomes less effective to apply the available research to outdoor environments.

The heat effect on thermal images has the worst outcome. It normally results in three different polarities of person representation in an image, as shown in Fig. 7.1: 1. person appearing darker as compared to the background. It occurs when the environmental temperature gets higher than the person temperature. 2. person appearing lighter than the background, which is when the person gets much warmer compared to the environment. 3. person appearance similar to the background. This happens when the person temperature as well as the environment temperature rise to the same level.

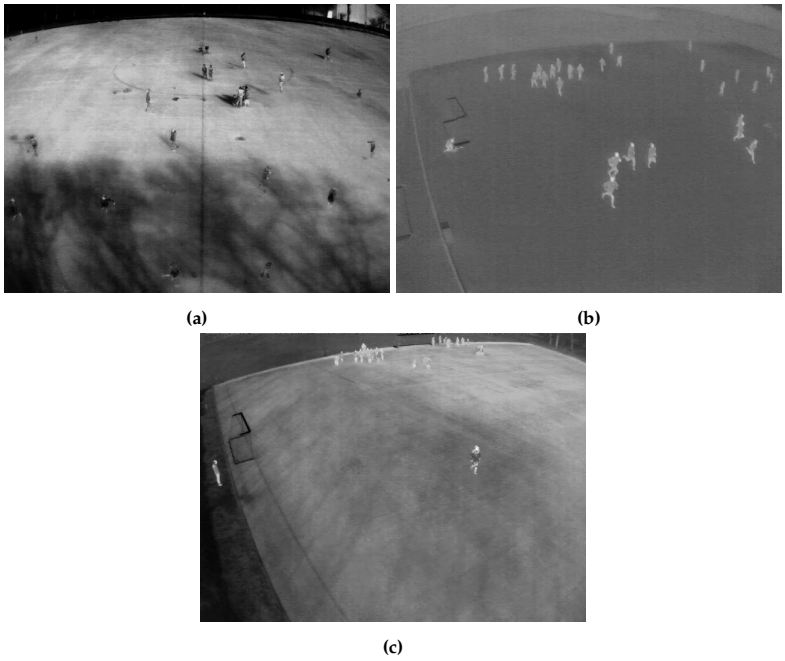



**Fig. 7.1:** Person representation in thermal images (a) Person appeared darker w.r.t background, (b) Person appeared lighter w.r.t background, (c) Person appeared similar to background

In this era of deep learning and big data solving, the approach to deal with the aforementioned or any other challenges is to increase the dataset either by producing synthetic data or by recording more videos. Increase the training data helps the classifier to learn every possible effect. On the other hand, some studies have proposed pre-processing techniques for thermal data to improve the performance of the convolutional neural network (CNN) [7, 8].

In this work, we have investigated the effects of pre-processing the thermal data for person detection using deep learning network. The implied pre-processing approach is to homogenize the data by converting the images to the same polarities/representation. Each representation is tested by using similar CNN and train settings to analyze which data type helps the network to perform better. We hypothesise that proposed pre-processing techniques should improve the performance of the subsequent algorithms. The pre-processing should help in model adaptation if they have an impact. Based on the above hypothesis following are the two main contributions in this paper,
1. Evaluation of polarity homogenization based pre-processing techniques

for person detection using a deep neural network in thermal images, 2. A new method for polarity detection and polarity homogenization in thermal data.

The rest of the paper is arranged as: Section 2 presents the overview of thermal person detection techniques as well as pre-processing methods explored both in thermal and RGB data. Section 3 describes the proposed experiments, and the results are presented in Section 4. Section 5 concludes this work.

# 2 Related work

In this section, deep learning solutions for thermal person detection and pre-processing techniques for enhancing deep network performance are introduced.

Many deep learning techniques have also been proposed for person detection. In [6] maximally stable extremal regions (MSERs) and CNN based methods were proposed for person detection using thermal pedestrian [1], OSU color thermal [2] and terravic motion IR [11] datasets. Tumas et. al [17] combined HOG and CNN for pedestrian detection for FIR domain. Heo et. al [5] combined YOLO and adaptive boolean-map-based saliency methods for pedestrian detection using CVC-09 [16] dataset. In [8] Huda et. al investigated the effects of environment and weather conditions for players detection in the outdoor soccer field. They used the transfer-learning approach using YOLO3 for analysis and evaluation. In [12], authors implemented a cascade object detector to detect human silhouette in thermal images. They aimed to develop a pedestrian detection system in poor light conditions. Zhang et. al [19] addressed the lack of availability of thermal dataset for implementing CNN networks. To deal with this challenge authors proposed RGB to thermal translation models to generate synthetic thermal data for training deep networks.

In the domain of deep learning several pre-processing techniques have been reported for improving the detection performance. In [13] zero component analysis is reported to have the most significant effect on the performance of image classification using CNN. The noise removing techniques i.e., non-local filtering, bilateral filtering and total variation denoising methods were studied to improve the image quality before it is passed to deep neural network [18]. Diah et. al [14] studied the influence of resizing, face detection, cropping, adding noise and normalization on CNN performance for emotion detection. Francisco et. al [10] found intensity normalization to have the most effect on the diagnosis of Parkinson's disease using CNN based

models. In [9] logarithmic and square root transformation methods were proposed for enhancing mammogram to detect breast cancer. Square root was found to have more influence on the performance of CNN based detection network. Image inversion, blurring, histogram stretching, and equalization methods were used to pre-process thermal images for person detection using CNN [7]. Homomorphic filtering and OTSU thresholding methods are proposed in [3] for improving image quality for concrete cracks detection using CNN. In [15] it is suggested that it is better to not pre-process the data. Their results showed that most CNN networks perform better if trained from scratch and only using data augmentation. Whereas, in [7] it is shown that inversion and histogram stretching techniques perform better while training a CNN based model using transfer learning for thermal person detection. Huda et. al [8] also proposed that inversion of thermal images to same person representation w.r.t background may help in the improvement of performance.

In the reported literature, we have perceived that the impact of pre-processing is mostly positive. In a few cases, the pre-processing does not improve the performance of the detection network. Our aim in this paper is to investigate and evaluate the impact of data homogenization based pre-processing technique using a CNN network performance on a thermal person detection dataset.

# 3 Methodology

In this work, the key investigation is to evaluate the role of image homogenization and image enhancement based pre-processing techniques used for transfer learning from a pretrained CNN network. We have used Yolov3, which is pretrained on RGB Imagenet data. The network is utilized due to its higher detection accuracy and the ability to detect smaller objects in an image.

As discussed earlier, heat effects alter the representation of a person in a thermal image in three ways.

- Person appeared lighter w.r.t background
- Person appeared darker w.r.t background
- Person appeared similar to background

To add more clarity in the images of similar background, we are using histogram stretching to enhance the person intensities in the background.

After histogram stretching, we have proposed and tested the following possible ways of passing the data to the learning network. The possibilities are graphically explained in Fig. 7.2 and the details are as follows.

- Normal data: Both train and test datasets are kept in their original form.
- Enhanced data: Histogram stretching is applied to both train and test datasets to enhance the image contrast.
- Light person: Train dataset is homogenized and enhanced. The homogenization is performed by inverting and making the person representation lighter on dark background for all the images. Image enhancement is performed by histogram stretching. The test dataset is altered by detecting and inverting the events of dark person representation on the lighter background (the procedure of detection is explained later in the section). The test data is also enhanced by using histogram stretching.
- Dark person: Train dataset is homogenized and enhanced. The homogenization is performed by inverting and making the person representation darker w.r.t the lighter background. Image enhancement is performed by histogram stretching. In the test dataset, the events of light person representation with a darker background are detected and inverted. The test data is also enhanced by using histogram stretching.
- Light person test on normal data: Train dataset is homogenized and enhanced. The homogenization is performed by inverting and making the person representation lighter on dark background for all the images. Image enhancement is performed by histogram stretching. Test data is used as it is.
- Dark person test on normal data: The homogenization is performed by inverting and making the person representation darker on light background for all the images. Image enhancement is performed by histogram stretching. Test data is used as it is.

For the testing data, the first step is the detection of the polarity of the image. Detection of events is performed in two steps, i.e., 1- sunlight detection and 2- human temperature detection.

**Sun light detection:** In thermal imagery, the images captured in high sunlight are the images that are brighter than the other images. For the detection of brighter images, image segmentation is performed. Sum of entropy-based thresholding is used for segmentation of images to get the brighter spots separated from the darker spots in the images. Afterwards, the accumulated pixel value is calculated by summing up all the pixels. The sum represents the number of bright pixels. A threshold is applied to the number of bright pixels to identify the images with high sunlight.

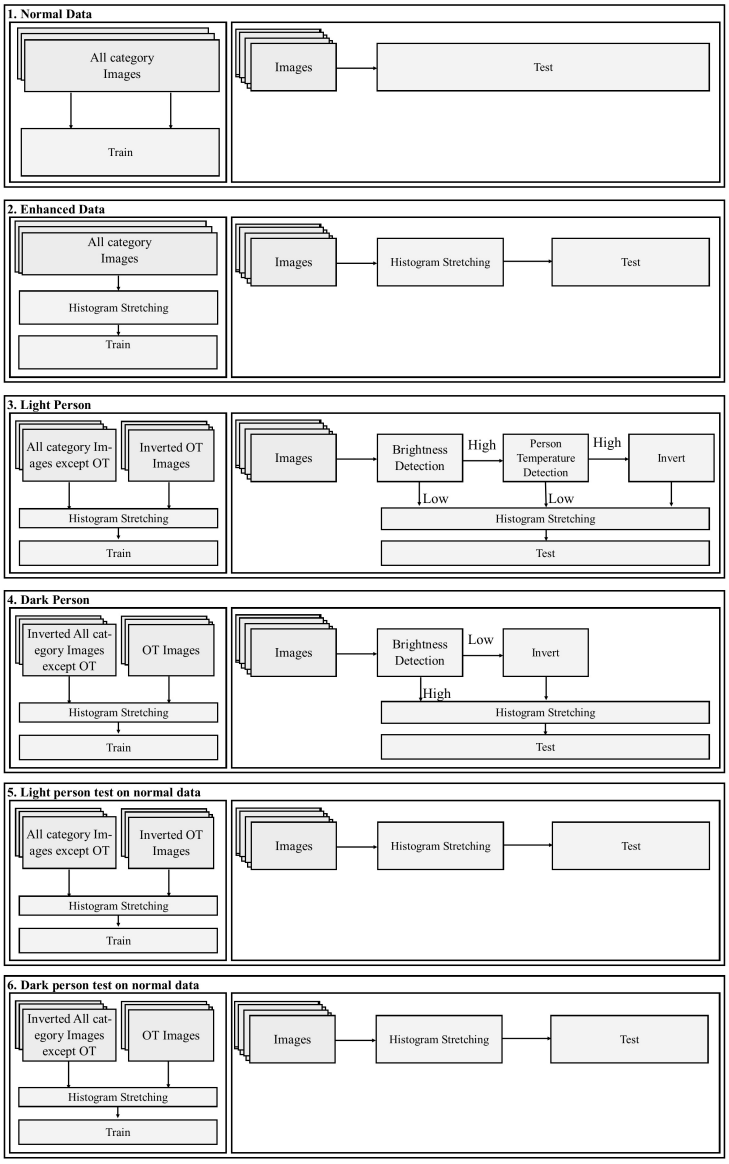



**Fig. 7.2:** Test setups

**Low human body temperature detection:** After the detection of brighter images, the next step is to find the lower person intensities in those brighter images. The data is separated based on histogram maps of both polarities of images i.e., light person with a dark background and dark person with a light background, separately. Two main features are used to detect low body temperature i.e., 1. intensity peaks vs the number of intensity bin in the histogram map of the image and 2. value of peak intensity in the histogram map.

# 4 Evaluation protocol

## 4.1 Training

### Dataset

For training, we have used the training data presented in [4, 8]. [4] is an indoor recorded soccer thermal dataset and consists of 3000 images. Every image consists of 8 persons playing soccer and occludes at some point while playing. The data is used as pre-training thermal data [8]. AAU-PD-T [8] training and testing data consist of 1941 and 1000 images, respectively, from the outdoor soccer field. In these images persons are running, playing, and doing exercises. The data is characterized as far players (far), images with snow (snow), images with the wind (wind), good outdoor (NR), occlusion (Oc), opposite temperature (OT), shadow (Sh), and similar temperature (ST).

### Network setting for training

Re-training of Yolov3 is performed using training data. In order to maintain consistency in results, training setup parameters, except number of iterations, are kept constant and exactly same as reported in [8] (i.e., learning rate = 0.001, momentum = 0.9, and decay = 0.0005). The iterations are set to 10000 to have long term analysis and observe deviations if any.

## 4.2 Testing

### Sun light detection

AAU-PD-T dataset is used for separating high-temperature images based on thresholding as defined in section III. In the AAU-PD-T dataset, OT and ST category contain the images with high sunlight. Fig. 7.3 clearly shows that sum of pixel values is directly related to image intensity values. NR category images also show some spikes for the sum of pixel values due to some images captured on sunny days.

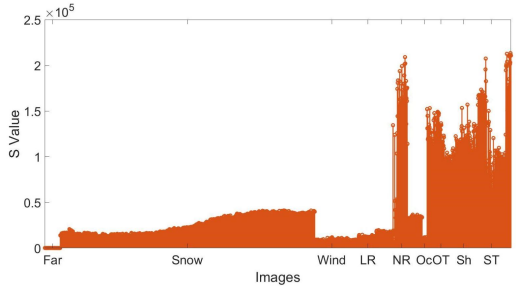



**Fig. 7.3:** Sum of pixel values vs categorical data

#### Low human body temperature detection

AAU-PD-T is used to observe the threshold values to separate the two polarities. It can be seen in fig. 7.4 that images with the person having lower body temperature than background seem to have lower intensities and a dip between bins 100 and 200. Also, these images mostly have two peaks. Whereas the images with persons having higher body temperature than background have high-intensity values especially in middle bins i.e., between bin 100 and bin 200.

#### Dataset for evaluating the testing scenerios

AAU-PD-T [8] test data and OSU-T [1] data are used for the evaluation of testing scenarios presented in Fig. 7.2. AAU-PD-T contains 1000 image with variable polarities. OSU-T dataset consists of 284 images. It is recorded in the Ohio State University campus at a pedestrian intersection and has images with both polarities.

## 4.3 Evaluation parameters

Precision and recall are used as evaluation parameters.

$$\text{F1 score} = \frac{2\text{xPrecisionxRecall}}{\text{Precision+Recall}} \tag{7.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \tag{7.2}$$

Here TP (True Positives) are the number of persons correctly detected, FP (False Positives) are the number of persons that are falsely detected. TN (True Negatives) are the correctly identified negative frames (frames with no persons) and FN (False Negatives) are the persons that are not detected. The training and testing of all combinations are performed using a graphical processing unit GTX 1080 with Linux Ubuntu 16.04.

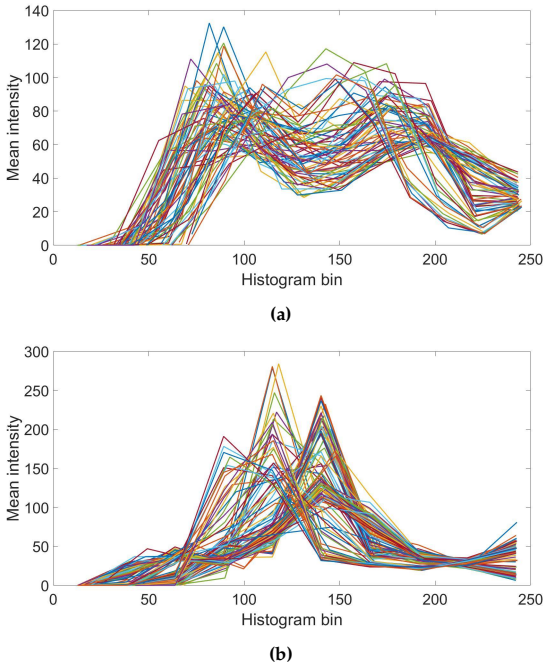
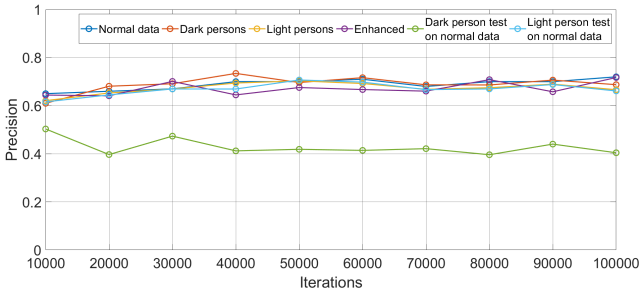
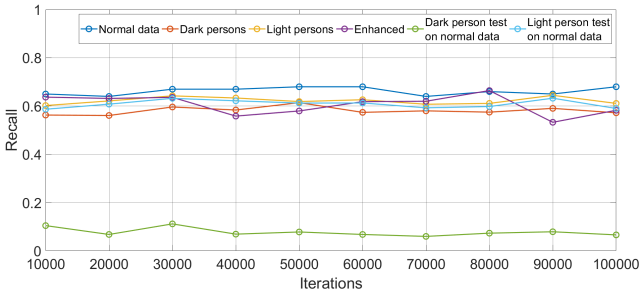



**Fig. 7.4:** Mean intensity values for respective histogram bins. (a) Images with lower body temperature than background. (b) Images with higher body temperature than background.




**(a)**




**(b)**

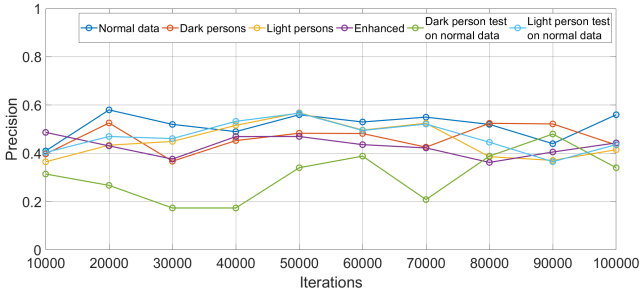
**Fig. 7.5:** Results of AAU-PD-T [8] dataset.

# 5 Results

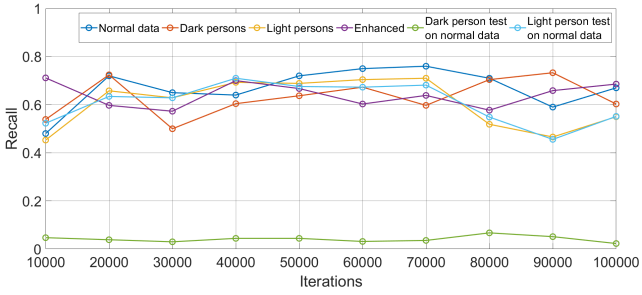
The results for each iteration are plotted in Fig. 7.5, and Fig. 7.6 and the results of max F1 Score for all iterations is presented in Table 7.1.

It can be seen from Fig. 7.5 and Fig. 7.6 that there is no significant difference in performance, i.e., precision and recall, between each testing setup except for dark person test on normal data. The performance of this setup is the lowest, which is understandable because most of the images in the datasets has a person representation of lighter intensity w.r.t the background. The results of the AAU-PD-T dataset shows that precision is similar for each test setup, whereas recall is higher for normal data test setup. The lower performance of recall for pre-processed datasets can be explained by the fact that training CNN network with high contrast images increases the chances of miss-detection for slightly lower contrast images.

The results of the OSU dataset shows that recall is higher than precision, which shows that FP detection is higher than FN detections. This is because the OSU dataset has better contrast images, and the person is labelled when appeared more than 50%. In the training, dataset person visibility varies and




**(a)**




**(b)**

**Fig. 7.6:** Results of OSU-T [1] dataset

it is not fixed to any percentage for labelling. Therefore, any person that is detected in the OSU dataset with visibility lower than 50% will appear as FP. This effect got enhanced when the images are pre-processed. The overall

**Table 7.1:** F1 Score for test setups

| Data | AAU-PD-T | OSU-T |
|---|---|---|
| Normal data | **0.70** | **0.65** |
| Enhanced data | 0.69 | 0.58 |
| Light person | 0.67 | 0.63 |
| Dark person | 0.65 | 0.61 |
| Light person test on normal data | 0.66 | 0.63 |
| Dark person test on normal data | 0.18 | 0.12 |

results in Table 7.1 show that the best performance is achieved when CNN is trained with normal data in its original form. When the network is trained with pre-processed data, it gets sensitive to a particular type of data and can react to any small change in testing data. In real-life applications, this kind of processed data can have more drastic effects as data is unpredictable.

# 6 Conclusion

In this work, the performance of a deep learning-based person detection network is analyzed for thermal data. The impact of inversion for creating homogeneity in person representation and histogram stretching for image enhancement were evaluated by proposing six testing setups. Results showed that the performance of the CNN network does not improve by pre-processing. Techniques improving the contrast of the images have a negative impact as the data in real-life scenarios is not restricted to better contrast images and increasing contrast may also enhance the noise. As different studies [7] have shown, pre-processing to improve data diversity and amount of data to process may help to improve detection performance. On the other hand, if the dataset is diverse and pre-processing is used to restrict the data in one form or to induces homogeneity, it causes degradation in the performance of the deep neural network.

# References

[1] James W Davis and Mark A Keck. A two-stage template approach to person detection in thermal imagery. In *IEEE Workshops on Applications of Computer Vision*, volume 1, pages 364–369, Jan. 2005.

[2] James W. Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2):162–182, 2007.

[3] Ronghua Fu, Hao Xu, Zijian Wang, Lei Shen, Maosen Cao, Tongwei Liu, and Drahomír Novák. Enhanced intelligent identification of concrete cracks using multi-layered image preprocessing-aided convolutional neural networks. *Sensors*, 20(7):2021, 2020.

[4] Rikke Gade and Thomas B. Moeslund. Constrained multi-target tracking for team sports activities. *IPSJ Transactions on Computer Vision and Applications*, 2018.

[5] Duyoung Heo, Eunju Lee, and Byoung Chul Ko. Pedestrian detection at night using deep neural networks and saliency maps. *Electronic Imaging*, 2018(17):060403–1–060403–9, 2018.

[6] Christian Herrmann, Thomas Müller, Dieter Willersinn, and Jürgen Beyerer. Real-time person detection in low-resolution thermal infrared imagery with mser and cnns. volume 9987, 2016.

[7] Christian Herrmann, Miriam Ruf, and Jürgen Beyerer. Cnn-based thermal infrared person detection by domain adaptation. In *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, volume 10643, page 1064308. International Society for Optics and Photonics, 2018.

[8] Noor Ul Huda, Bolette D Hansen, Rikke Gade, and Thomas B Moeslund. The effect of a diverse dataset for transfer learning in thermal person detection. *Sensors*, 20(7):1982, 2020.

[9] Agnese Marchesi, Alessandro Bria, Claudio Marrocco, Mario Molinara, Jan-Jurre Mordang, Francesco Tortorella, and Nico Karssemeijer. The effect of mammogram preprocessing on microcalcification detection with convolutional neural networks. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 207–212. IEEE, 2017.

[10] Francisco J Martinez-Murcia, Juan M Górriz, Javier Ramírez, and Andres Ortiz. Convolutional neural networks for neuroimaging in parkinson's disease: is preprocessing needed? *International journal of neural systems*, 28(10):1850035, 2018.

[11] R. Miezianko and D. Pokrajac. People detection in low resolution infrared videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–6, US, Jun. 2008.

[12] Adam Nowosielski, Krzysztof Małecki, Pawel Forczmański, Anton Smoliński, and Kazimierz Krzywicki. Embedded night-vision system for pedestrian detection. *IEEE Sensors Journal*, 2020.

[13] Kuntal Kumar Pal and KS Sudeep. Preprocessing for image classification by convolutional neural networks. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 1778–1781. IEEE, 2016.

[14] Diah Anggraeni Pitaloka, Ajeng Wulandari, T Basaruddin, and Dewi Yanti Liliana. Enhancing cnn with preprocessing stage in automatic emotion recognition. *Procedia computer science*, 116:523–529, 2017.

[15] Larissa Ferreira Rodrigues, Murilo Coelho Naldi, and João Fernando Mari. Comparing convolutional neural networks and preprocessing techniques for hep-2 cell classification in immunofluorescence images. *Computers in Biology and Medicine*, 116:103542, 2020.

[16] Yainuvis Socarras, Sebastian Ramos, David Vázquez, Antonio López, and Theo Gevers. Adapting pedestrian detection from synthetic to far infrared images. In *International Conference on Computer Vision (ICCV) Workshop*, jan. 2013.

[17] Paulius Tumas, Artūras Jonkus, and Artūras Serackis. Acceleration of hog based pedestrian detection in fir camera video stream. In *2018 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pages 1–4. IEEE, 2018.

[18] Jonghwa Yim and Kyung-Ah Sohn. Enhancing the performance of convolutional neural networks on quality degraded datasets. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2017.

[19] Lichao Zhang, Abel Gonzalez-Garcia, Joost van de Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 28(4):1837–1850, 2018.

# Part V

# Combinational Setup

# Chapter 8

# Multimodal and multiview distillation for real-time player detection on a football field

Anthony Cioppa, Adrien Deliège, Noor Ul Huda, Rikke Gade, Marc Van Droogenbroeck, Thomas B. Moeslund

The paper has been published in the
Conference on Computer Vision and Pattern Recognition Workshops
(CVPRW)

*The layout has been revised.*

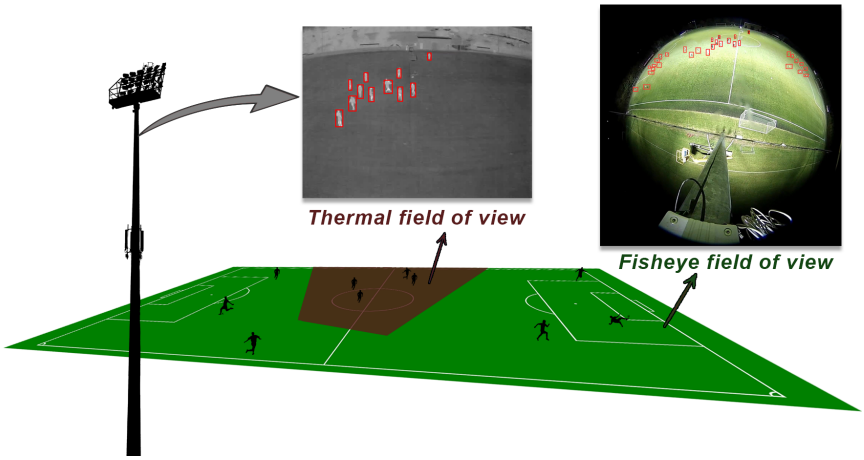



**Fig. 8.1:** Illustration of the problem handled in this paper. We leverage the detections made on a thermal image on a part of the field to detect all the players on the whole field on the fisheye image.


# Abstract

*Monitoring the occupancy of public sports facilities is essential to assess their use and to motivate their construction in new places. In the case of a football field, the area to cover is large, thus several regular cameras should be used, which makes the setup expensive and complex. As an alternative, we developed a system that detects players from a unique cheap and wide-angle fisheye camera assisted by a single narrow-angle thermal camera. In this work, we train a network in a knowledge distillation approach in which the student and the teacher have different modalities and a different view of the same scene. In particular, we design a custom data augmentation combined with a motion detection algorithm to handle the training in the region of the fisheye camera not covered by the thermal one. We show that our solution is effective in detecting players on the whole field filmed by the fisheye camera. We evaluate it quantitatively and qualitatively in the case of an online distillation, where the student detects players in real time while being continuously adapted to the latest video conditions.*


# 1 Introduction

Local sports fields can be expensive to construct and maintain, especially those built with artificial turf. Therefore, it is important to monitor and then optimize the occupancy of existing fields and stadiums. Furthermore, an automatic occupancy analysis method may open up new possibilities within

real-time information and booking. In this work we propose a robust and cost-effective method for player detection and counting in a football field.

For robust video monitoring of outdoor football fields, one main challenge is the size of the field. A field may be covered by either several regular cameras, which makes the setup rather complex and expensive, or it is possible to use a camera with a wide field of view, such as a fisheye camera. However, with a fisheye camera covering the entire football field, the players will appear small and have different orientation in the image due to the lens distortion. Player detection on these types of images is therefore not a trivial task. Another main challenge in outdoor environments is varying lighting conditions. Even though a football field may be illuminated during nights, lighting conditions will change during the day due to changing weather, position of the sun, and the effect of artificial lighting. To avoid problems with difficult lighting conditions, thermal cameras may be considered. These cameras capture only thermal infrared radiation, which represents temperature in the scene, hence they are more independent of lighting and normally eases the task of person detection because people have a temperature different from the background [14]. However, thermal cameras are expensive and due to their limited field of view and resolution, several cameras would be needed to cover a football field.

To construct a camera setup that is reasonable in price level and at the same time robust to changes in weather and lighting conditions, we propose to use one fisheye RGB and one thermal camera co-located at the side of the field. An illustration of the setup and example images from the two cameras are shown in Figure 8.1. Only the fisheye camera will cover the entire field, while the detections obtained directly from the thermal camera will serve to provide some kind of ground truth for teaching a network.

There are two main contributions in this paper: (i) We show how two different image modalities and fields of view can be combined in a student-teacher distillation approach. (ii) We show how a student network can be trained to detect players outside the field of view of the teacher, through a combination of a custom data augmentation process and a motion detection algorithm.

## 2 Related work

**Player detection in sports.** Detection of players in sports fields is the first step of vision systems for sports applications, like occupancy analysis, tracking, performance analysis, etc. [36]. Background subtraction based methods have often been used for player detection due to the fast processing time that makes it well-suited for real-time applications. It has been applied for static cameras [1, 33] and for moving cameras in the case of uniformly colored

surfaces [31]. However, noise should be expected due to, e.g., other moving objects, similar colors in foreground and background, changing lighting conditions, and shadows. It has also been proposed to use classic person detection methods like using the AdaBoost algorithm for training a linear classifier with HOG features for detecting players in Australian Rules Football [11], or similarly with AdaBoost and Haar features for player detection in basketball [21] and baseball [26].

More recently, like for general object detection, CNN-based methods have also been the dominant trend for detecting sports players. In [34] a shallow CNN was trained to detect players on a hockey field, while others use pretrained networks like Mask R-CNN for handball videos [30] and basketball videos [41], or YOLO for handball videos [6]. In [43] a reverse connected convolutional neural network (RC-CNN) is proposed for player detection. The reverse connected modules are embedded into the CNN to pass semantic information captured by deep layers back to shallower layers.

**Person detection in fisheye and thermal cameras.** Fisheye cameras have been widely used for person detection because of their advantage of wide viewing angle. Methods using a single camera setup have been reported for surveillance [22, 23], automobiles [24], indoor environment [35, 39] and outdoor sports field [18]. In these methods, the setup was used for pedestrian detection, tracking and occupancy analysis. Multiple camera setups are also proposed to detect persons for similar applications [3, 28, 40]. However, the main disadvantages with fisheye cameras are the distortion on the borders and the lower image quality in low lighting conditions.

Thermal cameras have long been used in practice because of their efficiency in bad lighting conditions. The range of applications varies from industrial uses to daily life traffic and surveillance [14]. Various methods based on thermal cameras have been proposed for person detection, such as feature extraction and threshold based methods [9, 12, 13, 42], HOG methods [25, 37], machine learning techniques [20] and deep neural networks [16, 17, 19]. A dataset and a trained network for people detection on outdoor thermal images have been proposed in [19]. The disadvantage of thermal cameras is their expensive cost and their reduced field of view.

In this work we will continue on recent trends to use a CNN-based method for player detection. We aim to circumvent the limitations of both fisheye and thermal cameras, by combining these modalities and teach the network for the fisheye camera with detections from the thermal camera, in a student-teacher distillation approach.

# 3 Data acquisition and calibration

**Camera setup.** The data used in this work consist of video streams of two different cameras: a fisheye camera and a thermal camera. Both cameras are installed on the same pole at the side of a football field, as illustrated in Figure 8.1. The thermal camera is placed approximately 9.8 meters above the ground and the fisheye camera is installed at 9.5 meters. By doing so, the field of view of the fisheye camera covers the whole football field, whereas the thermal camera covers the central area, as shown in Figure 8.1. In this setup, the field of view of the thermal camera represents 6% of the fisheye image, and covers 22% of the football field as seen by the fisheye camera.

Let us note that several teams use the field simultaneously for a training session during the video. Hence, the players are performing different activities, such as moving goals or performing various exercises. Therefore, the players can be found in different postures in any part of the field.

**Acquisition.** The fisheye video stream is recorded using a Hikvision Fisheye Network Camera with a resolution of $1280 \times 1280$ pixels and a field of view of $360^\circ$. The thermal video stream is recorded using an Axis Q1922 camera that has a resolution of $640 \times 480$ pixels and $57^\circ$ of horizontal viewing angle.

The videos were recorded during one hour in an amateur football field in December 2017, at night time with artificial light illuminating the field. The fisheye camera records the video at 12 fps. The thermal camera initially records the video at 30 fps, which is then re-sampled at 12 fps to allow a synchronization of the two streams. A proper camera calibration and registration between fisheye and thermal images is required for the transferability of points of interest.

**Calibration and registration.** First, a calibration of the internal parameters of each camera is performed following the procedure described in [29]. For the thermal camera, an A3-sized 10 mm polystyrene foam board is used as backdrop and a board of the same size with cut-out squares is used as checkerboard. In order to obtain a suitable contrast, the backdrop is heated and the checkerboard is placed at room temperature before the calibration. For the fisheye camera calibration, a checkerboard of $25 \times 25$ centimeters is used. Finally, the camera parameters derived from the calibration are obtained with a Matlab toolbox [4]. Second, we perform the registration between the two cameras. We undistord the images of the cameras using the internal parameters obtained previously. We manually choose several points of interest on the undistorded football field to compute the homography between the cameras, following [27]. These points are player feet positions for the players seen by the two cameras. The projection of the thermal image onto the fisheye image is shown in Figure 8.2.

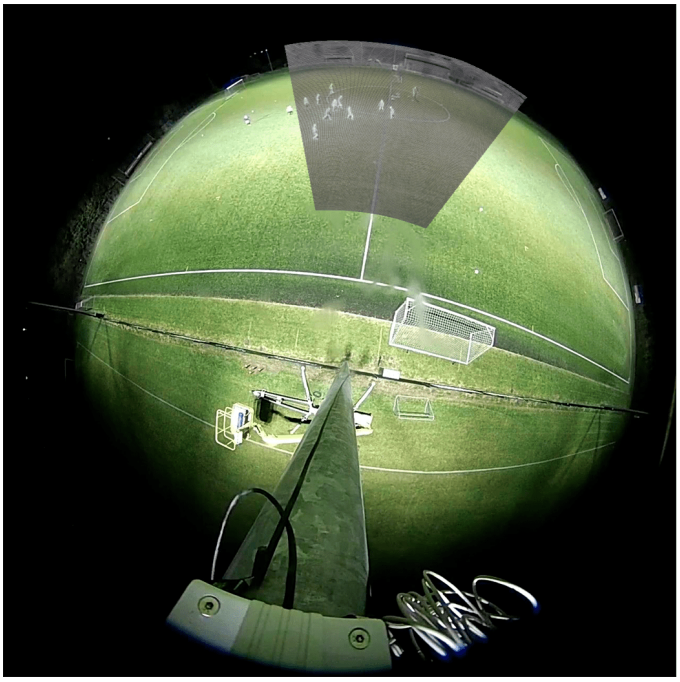


**Fig. 8.2:** Projection of the thermal image onto the fisheye image. The thermal camera sees only $\approx 22\%$ of the football field pixels of the fisheye image.

# 4 Methodology

**Problem statement.** A general formulation of the problem tackled in this paper is the following. Given a network performing a detection task on data from a camera, how can we train a real-time network for the same detection task on data from another camera with a possibly different modality and a different field of view of the same scene? In this section, we describe our solution for this problem in general terms, and we also explain how each step is particularized for our practical use case. Our use case consists in the task of player detection on a football field given a network able to detect players on a fixed thermal camera with a narrow field of view, which is used to train another detection network on data from a fixed fisheye camera with a wide field of view. This is illustrated in Figure 8.1.

**Notations.** We handle this problem with a teacher-student distillation approach, in which the output of a trained teacher network $\mathcal{T}$ erves as surrogate ground truth to train a student network $\mathcal{S}$ (see [38] for a recent review). Such a method has already been successfully applied in sports in [7] for segmenting football and basketball players in real time by distilling a slow $\mathcal{T}$ Mask R-CNN [15]) into a fast $\mathcal{S}$ (TinyNet [8]). In addition, in [7], the distillation is performed in an online fashion, such that $\mathcal{S}$ continuously adapts to the latest game conditions. However, $\mathcal{T}$ nd $\mathcal{S}$ use the same video feed, which implies that $\mathcal{S}$ can be directly (no transformation needed) and entirely (no missing ground truth) supervised by $\mathcal{T}$

In the present work, the setup is more challenging as $\mathcal{T}$ nd $\mathcal{S}$ process the video feeds of two cameras $\mathcal{C}_\mathcal{T}$ and $\mathcal{C}_\mathcal{S}$ with different modalities and fields of view. Having different modalities prevents us from using $\mathcal{T}$ n the feed of $\mathcal{C}_\mathcal{S}$, and having different fields of view prevents us from directly and entirely supervising $\mathcal{S}$. We assume that $\mathcal{C}_\mathcal{T}$ and $\mathcal{C}_\mathcal{S}$ are synchronized, such that they capture frames $\mathcal{C}_\mathcal{T}(t)$ and $\mathcal{C}_\mathcal{S}(t)$ simultaneously at each capture time $t$. We also assume that the projection from $\mathcal{C}_\mathcal{T}(t)$ to $\mathcal{C}_\mathcal{S}(t)$, expressed in terms of pixel coordinates, is known from the preliminary calibration step explained in the previous section. We note $\mathbb{P}$ the area of $\mathcal{C}_\mathcal{S}(t)$ representing the projection on $\mathcal{C}_\mathcal{S}(t)$ of the part of the scene also filmed by $\mathcal{C}_\mathcal{T}$ (shown in Figure 8.3). The remaining part of $\mathcal{C}_\mathcal{S}(t)$ is filmed by $\mathcal{C}_\mathcal{S}$ only and is noted $\overline{\mathbb{P}}$. As both cameras are fixed, this partition of $\mathcal{C}_\mathcal{S}(t)$ is independent of $t$.

In order to train $\mathcal{S}$, we need surrogate ground-truth bounding boxes both in $\mathbb{P}$ and in $\overline{\mathbb{P}}$. We detail hereafter how we obtain such boxes in $\mathcal{C}_\mathcal{S}(t)$ for a given capture time $t$. Following common practice, we represent a bounding box coordinates by a quadruplet containing the two coordinates of the center of the box, its width and its height.

**Surrogate ground truths in $\mathbb{P}$.** This part is straightforward. First, we use $\mathcal{T}$ o detect players in $\mathcal{C}_\mathcal{T}(t)$ and retrieve the coordinates of bounding boxes of

$\mathcal{C}_\mathcal{T}(t)$. Then, we project them into $\mathcal{C}_\mathcal{S}(t)$ using the calibration of the previous section. By doing so, we obtain the surrogate ground-truth bounding boxes of $\mathcal{C}_\mathcal{S}(t)$ that are located in $\mathbb{P}$, as shown in Figure 8.3. The remaining part of $\mathbb{P}$ constitutes detection-free areas.

**Surrogate ground truths in $\overline{\mathbb{P}}$.** This part is more difficult as we cannot have a direct access to the pixels of $\overline{\mathbb{P}}$ from those of $\mathcal{C}_\mathcal{T}(t)$. Training $\mathcal{S}$ solely with the boxes provided in $\mathbb{P}$ for each $\mathcal{C}_\mathcal{S}(t)$ leads the network to focus only on $\mathbb{P}$ and to overlook $\overline{\mathbb{P}}$ for each frame. Eventually, the network is not able to detect anything in $\overline{\mathbb{P}}$.

To circumvent this problem, our idea is the following. First, we use a custom data augmentation process to create artificial players with known bounding boxes in $\overline{\mathbb{P}}$. This provides us the "ground-truth locations" of some "true positive" players that $\mathcal{S}$ will have to detect. This is not sufficient as we still need "ground-truth information" in areas where we did not create any player. For that purpose, we use a motion detection algorithm to identify areas of $\overline{\mathbb{P}}$ that are guaranteed player-free. This provides us "true negative" areas, in which $\mathcal{S}$ will be penalized when predicting player bounding boxes. In the remaining areas of $\overline{\mathbb{P}}$, we have no useful information, hence $\mathcal{S}$ will not be penalized. These two steps are described in detail hereafter.

**[1. Custom data augmentation]** In order to introduce true positive players with known bounding boxes in $\overline{\mathbb{P}}$, we design the following automatic data augmentation process. Given a frame $\mathcal{C}_\mathcal{S}(t)$, we start by randomly extracting image crops delimited either by one isolated or by several adjacent bounding boxes previously obtained in $\mathbb{P}$ (Figure 8.4). Then, for each crop, we randomly select a pixel in $\overline{\mathbb{P}}$, which will serve as an anchor point where the crop will be pasted after being rescaled and rotated appropriately. In our use case, the anchors are selected in the subset of $\overline{\mathbb{P}}$ corresponding to the football field.

We perform a rescaling and a rotation of each crop to produce an insertion that looks as realistic as possible by taking into account the inherent distortions of $\mathcal{C}_\mathcal{S}$ (Figure 8.4). Let $(r, \theta)$ denote the initial polar coordinates (with origin located at the center of $\mathcal{C}_\mathcal{S}(t)$) of the center of the crop and $(r', \theta')$ those of its selected anchor point. We rescale the crop by a factor $\alpha e^{\beta(r'-r)} + \gamma$ with $\alpha = 0.5, \beta = -0.004, \gamma = 0.5$ and rotate it by the angle difference $\theta' - \theta$. Finally, we paste the transformed crop on $\mathcal{C}_\mathcal{S}(t)$ itself with OpenCV's seamless blending function, such that its center is located at the selected anchor point (Figure 8.4). In order to obtain the boxes associated with these artificial players, we perform the same transformation on each bounding box included in the initial crop. Eventually, for each transformed box, we consider as surrogate ground-truth bounding box the smallest unrotated (regular) rectangular box that encloses it (Figure 8.4).

In our fisheye setup, the data augmentation process allows to create artificial players with known bounding boxes in $\overline{\mathbb{P}}$ (Figure 8.4). However, this

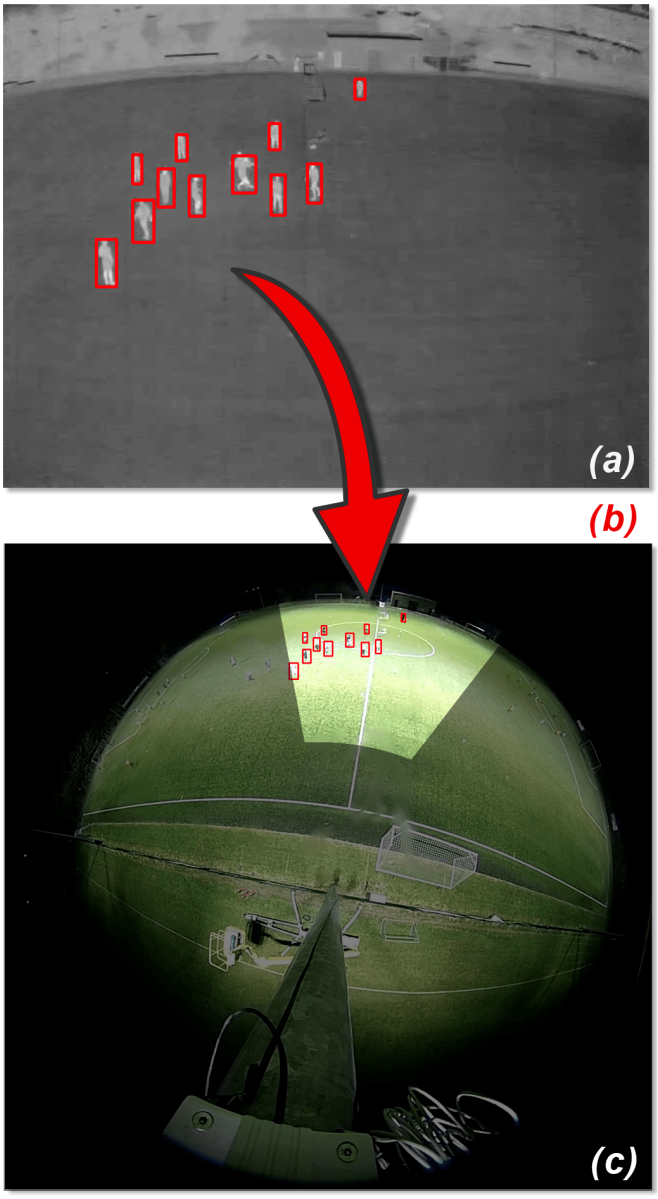



**Fig. 8.3:** The bounding boxes given by $\mathcal{T}$ n $\mathcal{C}_{\mathcal{T}}(t)$ (a) are projected (b) into $\mathcal{C}_{\mathcal{S}}(t)$ to provide us surrogate ground-truth bounding boxes in $\mathbb{P}$ (c).

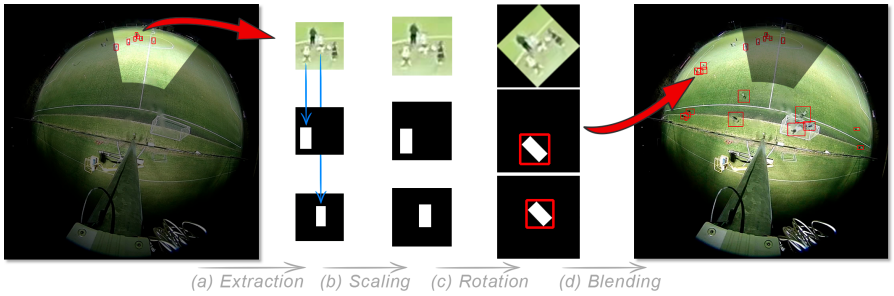



**Fig. 8.4:** Our custom data augmentation pipeline designed to construct surrogate ground-truth bounding boxes in the region $\overline{\mathbb{P}}$ filmed by $\mathcal{C}_{\mathcal{S}}$ only. First, crops containing players are extracted (a) from the area filmed by both cameras $\mathbb{P}$, in which we know their location. Then, each crop and its associated bounding boxes are scaled (b) and rotated (c) to be appropriately pasted in $\overline{\mathbb{P}}$. A seamless blending is applied during the collage to increase the realistic aspect of the augmented image. As a result, we create artificial players with known bounding boxes in $\overline{\mathbb{P}}$.

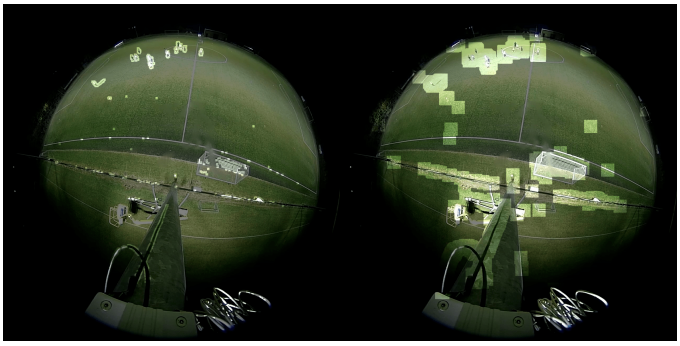


**Fig. 8.5:** Initial motion detection mask $\mathrm{M}(t)$ overlayed on its corresponding frame (left), and enlarged motion detection mask $\mathbb{M}(t)$ (right).

does not suffice to train $\mathcal{S}$ efficiently, as real players without known boxes may still be present in $\overline{\mathbb{P}}$. In a standard training process, $\mathcal{S}$ would thus be forced to detect the artificial players and would be penalized for detecting the remaining real ones. To bypass this undesirable effect, we remove the penalty suffered by $\mathcal{S}$ for detections containing enough motion. Hence, we leverage a motion detection algorithm to determine where this should be applied. By doing so, we also obtain areas where there is assuredly no player, i.e. where detections should not be made.

**[2. Motion detection]** As we handle a video feed from a fixed camera, we use ViBe [2] to obtain, for each frame $\mathcal{C}_{\mathcal{S}}(t)$, the set of pixels that are in motion, noted $\mathrm{M}(t)$, and those that are not, noted $\overline{\mathrm{M}(t)}$ (Figure 8.5). ViBe is very sensitive to motion, which implies that, in our fisheye setup, $\mathrm{M}(t)$ almost surely contains all the players, as well as pixels corresponding to the

balls, player shadows, and some noise. As $\mathbb{M}(t)$ may be tight around the players, we morphologically dilate it by a $11 \times 11$ square kernel to ensure that it includes the bounding boxes that would surround the players if they were available (Figure 8.5). By doing so, we obtain an enlarged mask $\mathbb{M}(t)$, such that we can penalize $\mathcal{S}$ when it detects players in $\overline{\mathbb{M}(t)}$, i.e. outside the enlarged mask. However, $\mathbb{M}(t)$ remains an area of uncertainty, where we do not penalize $\mathcal{S}$. Technically, this means that we zero out the loss in this area during training, as detailed hereafter.

**Training $\mathcal{S}$.** We use the YOLOv3 network [32] trained to detect humans on thermal images in [19] as teacher network $\mathcal{T}$ We use YOLOv3-tiny [32] as student network $\mathcal{S}$, adapted for a single class problem and with four times less channels for each convolutional layer. Hence, $\mathcal{S}$ outputs a list of 5-dimensional vectors. Each of them encapsulates information on a predicted bounding box: the four coordinates $(x, y, w, h)$ defining the box, and a player score $p$ representing its confidence for a player to actually belong to the box.

The loss of YOLOv3-tiny, hence $\mathcal{S}$, penalizes these vectors in the following way (see Figure 8.6). For a predicted box close to a surrogate ground-truth box (either in $\mathbb{P}$ or in $\overline{\mathbb{P}}$), the mean square error loss between the coordinates of the boxes is computed, as well as the binary cross-entropy loss of $p$. This encourages the network to predict a high confidence score (closer to 1) and to find the right dimensions of the box. For a box far from a surrogate ground-truth box, only the binary cross-entropy loss of $1 - p$ is computed, to discourage the network from predicting a player in that box ($p$ closer to 0). In our case, we must take into account the uncertainty about the boxes in $\mathbb{M}(t)$ in the region $\overline{\mathbb{P}}$, as they may correspond to unnanotated real players. Therefore, for a box far from a surrogate ground-truth box (including those created by the data augmentation), we zero out its loss if the center of the box is in $\overline{\mathbb{P}}$ and is in motion (belongs to $\mathbb{M}(t)$). If the center of the box belongs to $\overline{\mathbb{M}(t)}$, we are practically sure that there is no player in the box, and we thus leave the loss as is to penalize that detection. There is not particular restriction about the loss in $\mathbb{P}$. This is illustrated in Figure 8.6.

**Inference.** When used for inference, we verify that the bounding boxes predicted by $\mathcal{S}$ contain enough motion. Indeed, the predicted boxes whose center is not in motion, i.e. outside $\mathbb{M}(t)$, are not likely to contain a player. Therefore they are removed from the final output of $\mathcal{S}$.

# 5 Experiments

**Online distillation.** In this work, we perform the distillation of the teacher network $\mathcal{T}$ nto the student network $\mathcal{S}$ in an online manner as in [7]. The reason for using that process is threefold. First, this allows $\mathcal{S}$ to continuously

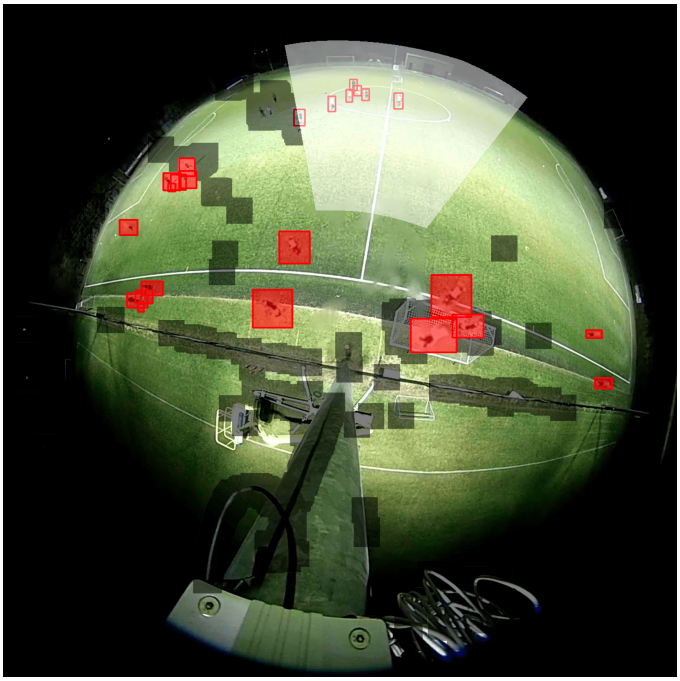


**Fig. 8.6:** Combination of our data augmentation and motion detection algorithms, showing how the loss is applied to penalize the predictions of $\mathcal{S}$ in $\overline{\mathbb{P}}$ (outside the white area). $\mathcal{S}$ must detect the players artificially created (red rectangles). Also, predicted boxes whose center falls within the enlarged motion mask $\mathbb{M}(t)$ (the black zones) do not generate any loss, since this area includes the players of $\overline{\mathbb{P}}$ not erased by the data augmentation, for which we have no ground-truth boxes. Finally, $\mathcal{S}$ must not predict any box in the rest of the image in $\overline{\mathbb{P}}$. Let us recall that the loss is applied everywhere in $\mathbb{P}$, as we have the ground truth from $\mathcal{T}$ n that area.

adapt to the latest weather and lighting conditions. Second, in a real-life deployment of the system, the online distillation will indeed be performed continuously. Hence, in order to have an understanding of how $\mathcal{S}$ behaves as it trains and detects people in real time, it is worth testing $\mathcal{S}$ under similar conditions. Third, training $\mathcal{S}$ adaptively allows us to study the evolution of the performance of the network as it learns through time. As we have only one video sequence with both the thermal and the fisheye recordings, this also enables us to evaluate $\mathcal{S}$ multiple times rather than measuring its performance only once, on a unique (and maybe abnormally hard or easy) small set of frames.

In the online distillation process, all the frames of the fisheye camera $\mathcal{C}_\mathcal{S}$ are treated by $\mathcal{S}$, which runs in real time. Meanwhile, some frames of the video feed of the thermal camera $\mathcal{C}_\mathcal{T}$ are input to $\mathcal{T}$ which provides boxes converted into surrogate ground-truth bounding boxes in the area $\mathbb{P}$ of the frame captured by $\mathcal{C}_\mathcal{S}$. These boxes are accumulated in an online dataset with 5-minutes memory, and the dataset is used to train a copy of $\mathcal{S}$ in a separate thread. The training is performed on the whole frames $\mathcal{C}_\mathcal{S}(t)$ as described in the previous section, using our data augmentation and motion detection processes outside $\mathbb{P}$. When this copy of $\mathcal{S}$ has trained during one epoch on the online dataset, its weights are updated and transferred into the initial network $\mathcal{S}$ that performs the detection on all the frames. Consequently, the weights of this network evolves through time to continuously adapt to the latest video conditions.

**Quantitative evaluation.** To assess the performance of the student network $\mathcal{S}$ over the course of the video, we manually annotated the ground-truth bounding boxes for all the players of one frame every 10 seconds of the fisheye video. We compute the performance of $\mathcal{S}$ on a set of frames with the Average Precision (AP) metric particularized for one class. Following practice for the Pascal VOC dataset [10], each bounding box predicted by $\mathcal{S}$ is matched with the ground-truth box with which it has the largest intersection over union (IoU). We consider predicted boxes with an IoU larger than some threshold t_IoU as true positives, the others as false positives, and the ground-truth boxes left unmatched are false negatives. If several true positives are associated with the same ground-truth box, only one of them is kept as a true positive, while the others are rather considered as false positives. We note the number of true positives (resp. false positives, false negatives) TP (resp. FP, FN). Then, we compute the precision and recall as

$$P = \frac{TP}{TP + FP} \quad \text{and} \quad R = \frac{TP}{TP + FN}.$$

We compute the points $(P, R)$ for various thresholds on the confidence scores of the boxes to obtain the PR curve. Finally, we compute the area under the PR curve as suggested in [10] to obtain the AP for that set of frames. Despite

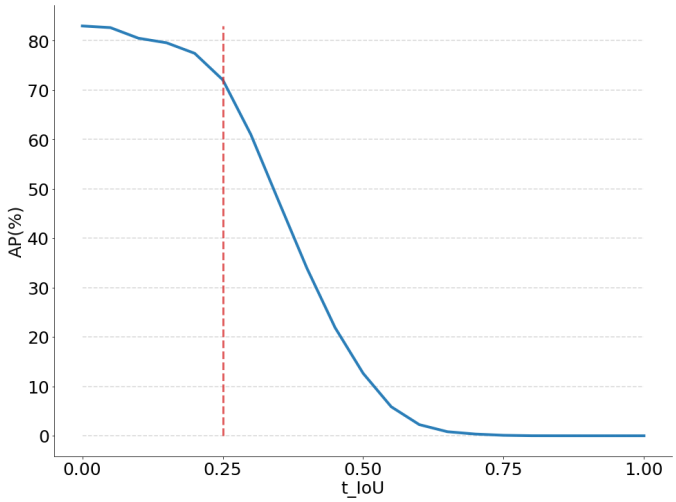



**Fig. 8.7:** **Performances of $\mathcal{T}$ n $\mathbb{P}$** on the last 15 minutes of video as a function of t_IoU. This quantifies how accurately $\mathcal{T}$ enters its bounding boxes on the players. We can see that $\mathcal{T}$ s not perfect. We decide to evaluate the **performances of $\mathcal{S}$ for t_IoU $= 0.25$**, as we consider it as the largest t_IoU for which $\mathcal{T}$ till displays satisfying performances (AP $> 70\%$).

its limitations [5], this kind of evaluation process has been widely adopted in the community.

In order to determine an appropriate value of t_IoU for evaluating the performance of $\mathcal{S}$, we examine the efficiency of $\mathcal{T}$ n predicting the boxes in $\mathbb{P}$. For that purpose, we compute the AP of $\mathcal{T}$ n the last 15 minutes of video, for several values of t_IoU ranging from 0 to 1, for the frames where ground-truth annotations are available. This allows us to determine how good $\mathcal{T}$ s at centering its bounding boxes on the players. The performance of $\mathcal{T}$ n $\mathbb{P}$ as a function of t_IoU is shown in Figure 8.7. We can see that $\mathcal{T}$ s not perfect in $\mathbb{P}$, which conditions the performances that can be expected from $\mathcal{S}$. To evaluate $\mathcal{S}$, we choose t_IoU $= 0.25$, as $\mathcal{T}$ isplays reasonable performances in $\mathbb{P}$ with that threshold. Given the small size of the boxes, it also makes sense to examine the performance of $\mathcal{S}$ for a relatively low value of t_IoU. Let us recall that the boxes outputted by the network are independent of any particular choice of threshold. It serves only for quantitative evaluation purposes.

Following [7], we evaluate the performance of the student network $\mathcal{S}$ progressively. Every 10 seconds, $\mathcal{S}$ predicts the bounding boxes of the frames for which we have manual annotations within a running temporal window that covers the next 3 minutes of video. For this set of frames, we compute the AP. The evolution on the AP through time with t_IoU $= 0.25$ is represented in Figure 8.8. We see that the performance tends to increase, indicating that

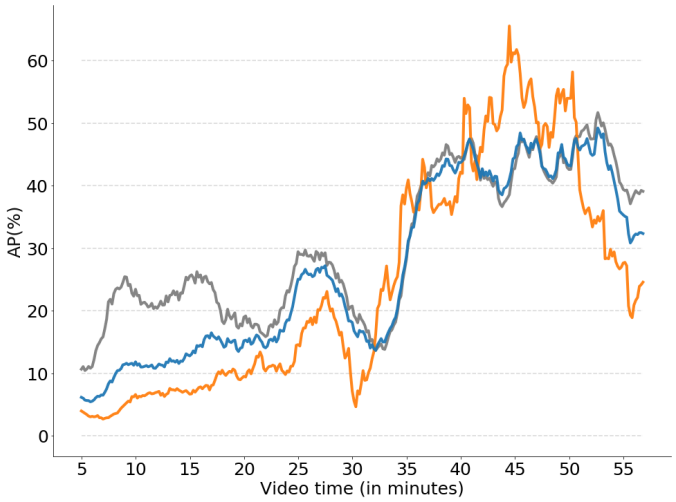



**Fig. 8.8:** Evolution of the performances of the student network $\mathcal{S}$ through the video in $\mathbb{P}$, $\overline{\mathbb{P}}$, and in the **whole frames**. We can see that the network improves over time and that it manages to perform well both in $\mathbb{P}$ and in $\overline{\mathbb{P}}$.

$\mathcal{S}$ learns to better detect players over time. Figure 8.8 also reveals that there is still room for improvement in the present challenge.

We further examine the effectiveness of our data augmentation and motion detection processes to train $\mathcal{S}$ for detecting players outside $\mathbb{P}$. For that purpose, we perform a region-specific analysis by computing the temporal evaluation of the AP within $\mathbb{P}$ and $\overline{\mathbb{P}}$. The performance curves are displayed in Figure 8.8. We note that $\mathcal{S}$ learns efficiently to detect players in $\overline{\mathbb{P}}$, as the performances for $\mathbb{P}$ and $\overline{\mathbb{P}}$ are close to each other and follow the same trend. Also, further experiments reveal that the post-processing with the motion mask $\mathbb{M}(t)$ is particularly helpful to increase the performance in $\overline{\mathbb{P}}$. In that area, the AP decreases by 5 to 20% without post-processing, while the drop is below 3% in $\mathbb{P}$.

Finally, as a potential application of this system is to monitor the use of the football field, we examine the results obtained for the task of people counting. The predicted number of people on the field corresponds to the number of bounding boxes predicted by $\mathcal{S}$ (thus on the fisheye images) after post-processing. We average the counting using a 1-minute sliding window. The results are displayed in Figure 8.9. We note that our method gives a globally reliable estimate of the number of people present on the field. Quantitatively, during the last 15 minutes of video, the root mean square error (RMSE) between the predictions and the ground truth is as low as 3.4 players. Again, we can see that the performance tends to increase over time since the estimate is more accurate at the end of the video, indicating that $\mathcal{S}$ learns to better

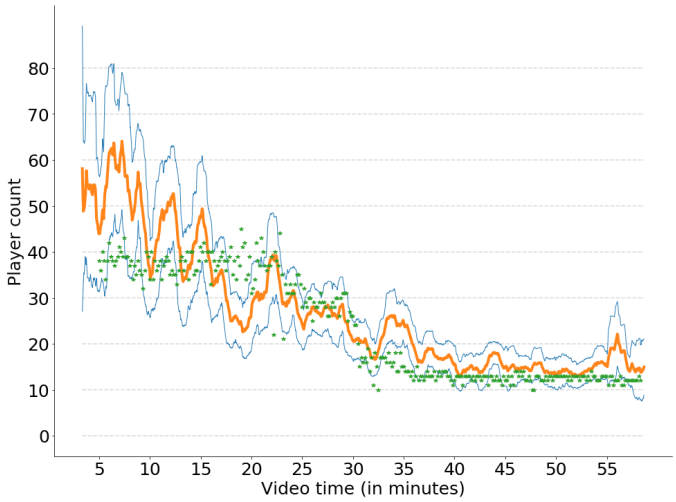



**Fig. 8.9:** **Results on the player counting task** averaged over a 1-minute window, and associated **standard deviation**. During the last 15 minutes, we have a RMSE with the **ground truth** of 3.4 players, which is reasonable and shows that our method provides a reliable estimate of the occupancy of the football field.

detect players over time. Also, we can see in Figure 8.9 that the standard deviation of the box count computed for each sliding window decreases over time, which indicates that the network becomes more consistent as it trains. Even though $\mathcal{S}$ tends to slightly overestimate the actual number of players, we can see that it manages to provides a good overview of the use of the field.

**Qualitative evaluation.** To further assess the usefulness of our data augmentation and motion detection processes, we perform ablation studies on the components of our method. We investigate the combination of either enabling or disabling the data augmentation, with either zeroing out the loss in the motion mask $\mathbb{M}(t)$, or nowhere in $\overline{\mathbb{P}}$, or everywhere in $\overline{\mathbb{P}}$. The effects observed for these setups are reported in Table 8.1. In our experiments, we observe that the combination of the data augmentation and of zeroing out the loss in $\mathbb{M}(t)$, as detailed in this paper, leads to the best student network $\mathcal{S}$ at inference time. Activating the loss everywhere in $\overline{\mathbb{P}}$ at training time forces $\mathcal{S}$ to detect only the artificial players in $\overline{\mathbb{P}}$ and to avoid detecting the actual players of $\overline{\mathbb{P}}$ that have not been erased by the data augmentation. This may confuse $\mathcal{S}$, leading to a decrease in its ability to detect players in $\overline{\mathbb{P}}$ at inference time. We notice that canceling the loss everywhere in $\overline{\mathbb{P}}$ leads to thousands of predicted bounding boxes in $\overline{\mathbb{P}}$ at inference time. This makes sense since the network is not forced to detect or not players in $\overline{\mathbb{P}}$ in this case. Most of these predictions are false positives, and the system is useless in

| In $\overline{\mathbb{P}}$ | With data augmentation | Without data augmentation |
|---|---|---|
| Cancel loss in the motion mask $\mathbb{M}(t)$ | **Our full method.** Most players in $\overline{\mathbb{P}}$ correctly detected, few false positives. | Few players detected in $\overline{\mathbb{P}}$, unusable in practice |
| Activate loss everywhere in $\overline{\mathbb{P}}$ | Able to detect players in $\overline{\mathbb{P}}$, but not as good as our full method | Unable to make any detection in $\overline{\mathbb{P}}$, no true positives |
| Cancel loss everywhere in $\overline{\mathbb{P}}$ | Thousands of detections in $\overline{\mathbb{P}}$, mostly false positives | Thousands of detections in $\overline{\mathbb{P}}$, mostly false positives |

**Table 8.1:** Ablation results in $\overline{\mathbb{P}}$. The combination of the data augmentation and the motion detection algorithm gives the best trade-off between true and false positive detections.

practice. As indicated in Table 8.1, we also note that removing the data augmentation always leads to mediocre networks, for similar reasons as those already explained. In particular, activating the loss everywhere in $\overline{\mathbb{P}}$ makes $\mathcal{S}$ unable to detect any single player in $\overline{\mathbb{P}}$. This results from the absence of ground-truth true positives (both artificial and real ones) in $\overline{\mathbb{P}}$.

Finally, examples of detections provided by $\mathcal{S}$ are given in Figure 8.10. We can see that players located in $\overline{\mathbb{P}}$ are detected as efficiently as those located in $\mathbb{P}$. This was made possible thanks to our data augmentation and motion detection algorithms in the distillation approach.

# 6 Conclusion

In this work, we propose a novel system for monitoring the field occupancy in low-budget football stadiums. Our system uses a single wide-angle fisheye camera assisted by a thermal camera to detect and count all the players on the field. We use a network trained in a student-teacher distillation approach. The student network is locally supervised by a teacher network that easily detects players on the thermal camera. These detections are then projected into the fisheye camera using camera registration and serve as surrogate ground truths. Since both cameras have different modalities and fields of view of the scene, the student cannot be fully supervised by the teacher. Therefore, we develop a custom data augmentation process, combined with motion information provided by a background subtraction algorithm, to introduce surrogate ground truths outside their common field of view. In our case, we perform the distillation in an online fashion, i.e. our student is continuously

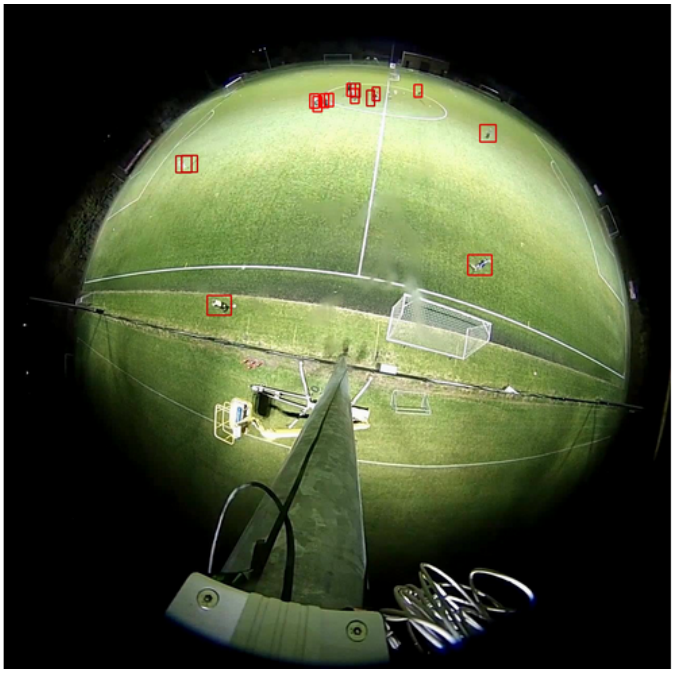


**Fig. 8.10:** Detections on a test frame. We can note that players are accurately detected, even though there are a few superfluous predicted bounding boxes.

trained to adapt to the latest video conditions, while performing the player detection in real-time. We show that our system is able to accurately detect players both inside and outside the common field of view, thanks to our custom supervision.

**Acknowledgments** A. Cioppa is funded by the FRIA. A. Deliège is supported by the DeepSport project of the Walloon region, Belgium.

# References

[1] M. Archana and M. Kalaiselvi Geetha. An efficient ball and player detection in broadcast tennis video. In *Intelligent Systems Technologies and Applications*, pages 427–436, Cham, 2016. Springer International Publishing.

[2] Olivier Barnich and Marc Van Droogenbroeck. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, June 2011.

[3] Massimo Bertozzi, Luca Castangia, Stefano Cattani, Antonio Prioletti, and Pietro Versari. 360° detection and tracking algorithm of both pedestrian and vehicle using fisheye images. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 132–137, June 2015.

[4] Jean-Yves Bouguet. Camera calibration toolbox for Matlab, 2014.

[5] Kendrick Boyd, Vítor Santos Costa, Jesse Davis, and C. David Page. Unachievable region in precision-recall space and its effect on empirical evaluation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML)*, ICML'12, page 1619–1626, Madison, WI, USA, 2012. Omnipress.

[6] Matija Buric, Marina Ivasic-Kos, and Miran Pobar. Player tracking in sports videos. In *IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 334–340, Dec. 2019.

[7] Anthony Cioppa, Adrien Deliège, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive Real-Time Human Segmentation in Sports Through Online Distillation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[8] Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games. In *IEEE Conference on Computer*

*Vision and Pattern Recognition (CVPR) Workshops*, pages 1846–1855, June 2018.

[9] Congxia Dai, Yunfei Zheng, and Xin Li. Layered representation for pedestrian detection and tracking in infrared imagery. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR) Workshops*, Sep. 2005.

[10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[11] Hayden Faulkner and Anthony Dick. AFL player detection and tracking. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, Nov. 2015.

[12] Rikke Gade, Anders Jørgensen, and Thomas B. Moeslund. Occupancy analysis of sports arenas using thermal imaging. In *International Conference on Computer Vision Theory and Applications*, pages 277–283. SCITEPRESS Digital Library, 2012.

[13] Rikke Gade, Anders Jørgensen, and Thomas B. Moeslund. Long-term occupancy analysis using graph-based optimisation in thermal imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3698–3705, US, 2013. IEEE Computer Society Press.

[14] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: A survey. *Machine Vision and Applications*, 25(1):245–262, 2014.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct. 2017.

[16] Duyoung Heo, Eunju Lee, and Byoung Chul Ko. Pedestrian detection at night using deep neural networks and saliency maps. *Journal of Imaging Science and Technology*, 61:60403–1–60403–9(9), 2017.

[17] Christian Herrmann, Thomas Müller, Dieter Willersinn, and Jürgen Beyerer. Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs. In *SPIE Security + Defence*, volume 9987, 2016.

[18] Noor Ul Huda, Bolette D. Hansen, Rikke Gade, and Thomas B. Moeslund. Occupancy analysis of soccer fields using wide-angle lens. In *International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pages 354–359, Dec. 2017.

[19] Noor Ul Huda, Bolette D. Hansen, Rikke Gade, and Thomas B. Moeslund. The effect of diverse dataset for transfer learning in thermal person detection. *Sensors*, 20(7):1982, Apr 2020.

[20] Noor Ul Huda, Kasper Halkjær Jensen, Rikke Gade, and Thomas B. Moeslund. Estimating the number of soccer players using simulation-based occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1937–1946, US, 2018. IEEE.

[21] Zdravko Ivankovic, Branko Markoski, Miodrag Ivkovic, Dragica Radosav, and Predrag Pecev. Adaboost in basketball player identification. In *IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 151–156, Nov. 2012.

[22] Hyungtae Kim, Eunjung Chae, Gwanghyun Jo, and Joonki Paik. Fisheye lens-based surveillance camera for wide field-of-view monitoring. In *IEEE International Conference on Consumer Electronics (ICCE)*, pages 505–506, 2015.

[23] Hyungtae Kim, Jaehoon Jung, and Joonki Paik. Fisheye lens camera based surveillance system for wide field of view monitoring. *Optik*, 127(14):5636–5646, 2016.

[24] Dan Levi and Shai Silberstein. Tracking and motion cues for rear-view pedestrian detection. In *IEEE International Conference on Intelligent Transportation Systems*, pages 664–671, Sep. 2015.

[25] Wei Li, Dequan Zheng, Tiejun Zhao, and Mengda Yang. An effective approach to pedestrian detection in thermal imagery. In *International Conference on Natural Computation*, pages 325–329, May 2012.

[26] Zahid Mahmood, Tauseef Ali, and Shahid Khattak. Automatic player detection and recognition in images using adaboost. In *International Bhurban Conference on Applied Sciences Technology (IBCAST)*, pages 64–69, Jan. 2012.

[27] Ezio Malis and Manuel Vargas. Deeper understanding of the homography decomposition for vision-based control. Research Report RR-6303, INRIA, 2007.

[28] Van Tuan Nguyen, Thanh Binh Nguyen, and Sun-Tae Chung. ConvNets and AGMM based real-time human detection under fisheye camera for embedded surveillance. In *International Conference on Information and Communication Technology Convergence (ICTC)*, pages 840–845. IEEE, 2016.

[29] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmose, Thomas B. Moeslund, and Sergio Escalera. Multi-modal RGB-depth-thermal human body segmentation. *International Journal of Computer Vision*, 118(2):217–239, 2016.

[30] Miran Pobar and Marina Ivasic-Kos. Mask R-CNN and optical flow based method for detection and marking of handball actions. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6, Oct. 2018.

[31] Upendra M. Rao and Umesh C. Pati. A novel algorithm for detection of soccer ball and player. In *International Conference on Communications and Signal Processing (ICCSP)*, pages 344–348, Apr. 2015.

[32] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[33] Vito Renò, Nicola Mosca, Massimiliano Nitti, Tiziana Dorazio, Donato Campagnoli, Andrea Prati, and Ettore Stella. Tennis player segmentation for semantic behavior analysis. In *IEEE International Conference on Computer Vision (ICCV) Workshop*, pages 718–725, Dec. 2015.

[34] Melike Şah and Cem Direkoğlu. Evaluation of image representations for player detection in field sports using convolutional neural networks. In *International Conference on Theory and Application of Fuzzy Systems and Soft Computing (ICAFS)*, pages 107–115, Cham, 2019. Springer International Publishing.

[35] Mamoru Saito, Katsuhisa Kitaguchi, Gun Kimura, and Masafumi Hashimoto. People detection and tracking from fish-eye image based on probabilistic appearance model. In *SICE Annual Conference 2011*, pages 435–440, Sep. 2011.

[36] 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.

[37] Paulius Tumas, Artūras Jonkus, and Artūras Serackis. Acceleration of HOG based pedestrian detection in FIR camera video stream. In *Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pages 1–4, Apr. 2018.

[38] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *CoRR*, 2020.

[39] Tsaipei Wang, Chia-Wei Chang, and Yu-Shan Wu. Template-based people detection using a single downward-viewing fisheye camera. In *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 719–723, Nov. 2017.

[40] Tsaipei Wang and Chih-Hao Liao. People detection in downward-viewing fisheye camera networks using fuzzy integral. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–5, June 2019.

[41] Yukun Yang, Min Xu, Wanneng Wu, Ruiheng Zhang, and Yu Peng. 3D multiview basketball players detection and localization based on probabilistic occupancy. In *Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, Dec. 2018.

[42] Hui Zhang, Baojun Zhao, Linbo Tang, Jianke Li, and Jianke Li. Variational-based contour tracking in infrared imagery. In *International Congress on Image and Signal Processing*, pages 1–5, Oct 2009.

[43] Lijing Zhang, Yao Lu, Ge Song, and Hanfeng Zheng. RC-CNN: Reverse connected convolutional neural network for accurate player detection. In *Pacific Rim International Conference on Artificial Intelligence (PRICAI): Trends in Artificial Intelligence*, pages 438–446, Cham, 2018. Springer International Publishing.

# Part VI

# Conclusion

# Chapter 9

# Report to municipality

This thesis work was supported by Aalborg Municipality, where the aim was to investigate and develop an autonomous system for occupancy analysis. This chapter will summarize the report that was prepared for the Municipality. The data presented in this report is from the studies conducted in this thesis.

## 1 Setup

The deployed camera setup adopted and proposed for long-term occupancy analysis is from Chapters 5 and 6, i.e. three thermal camera setup to cover the whole field. The camera setup is employed because it is independent of day/night light conditions.

The cameras are placed almost 10m from the ground and attached to a pole. As the fields are amateurs, camera placements vary depending on the pole's availability and electricity resources. The arrangements with respect to each soccer field are described in the coming sections.

## 2 Data recordings

The data is recorded from 10 different artificial grass fields from the 10th of January to the 11th of April, 2018. The recording days in each soccer field are described in Table 9.1.

The recording is performed at one $1/8$ $frame/sec$to meet up with long-term storage. Afterwards, data is prepared for further processing by performing time synchronization and image segmentation. First, the cameras are clocked for time synchronization for multiple cameras covering one field. The image segmentation is performed for overlapping and non-field areas

| Start date | End date | Name of the field |
|---|---|---|
| 10th of January | 4th of February | **1.** NFB-11 (Nørresundby Forenede Boldklubber, 11-players)<br>**2.** NFB-8 (Nørresundby Forenede Boldklubber, 8-players) |
| 8th of February | 28th of February | **1.** Aalborg Chang<br>**2.** Gug Boldklub |
| 5th of March | 18th of March | **1.** SSB (Storvorde Sejlflod Boldklub)<br>**2.** AAB-1 (Aalborg Boldspilklub af 1885, kunstgræsbane 1)<br>**3.** AAB-2 (Aalborg Boldspilklub af 1885, kunstgræsbane 2) |
| 21st of March | 11th of April | **1.** IAF (Idrætsklubben Aalborg Freja)<br>**2.** SGI-1 (Svenstrup-Godthåb Idrætsforening, 11-mands)<br>**3.** SGI-2 (Svenstrup-Godthåb Idrætsforening, 4 små |

**Table 9.1:** Overview of the recording periods for each soccer field.

by manual labels using a mapping video clip. The mapping video clip is recorded for every soccer field before the start of the actual recording. In that mapping video, a dummy person takes a round trip around the field. The video annotator follows the person in the video clip to map the field area and overlap of field areas in each camera. The overlapping and non-field areas are then segmented out from actual images of the video.

For the six fields named NFB-8, NFB-11, Aalborg Chang, Gug Boldklub, IAF and SSB, full field coverage is performed using three cameras to cover the whole field. Fig. 9.1 shows the camera coverage, and Fig. 9.2 shows the example images.

At fields AAB-1 and AAB-2, three cameras are installed to cover both fields, where both soccer fields share the middle camera. Camera coverage is demonstrated in Fig. 9.3, and example images are shown in Fig. 9.4.

Coverage of SGI-1 and SGI-2 is performed by using only one camera. Both fields are closely constructed, where SGI-1 is a big field, and alongside, there are four small fields. SGI-2 represents those four small fields. The camera placement and the soccer fields are explained in Fig. 9.5, and the example images are shown in Fig. 9.6

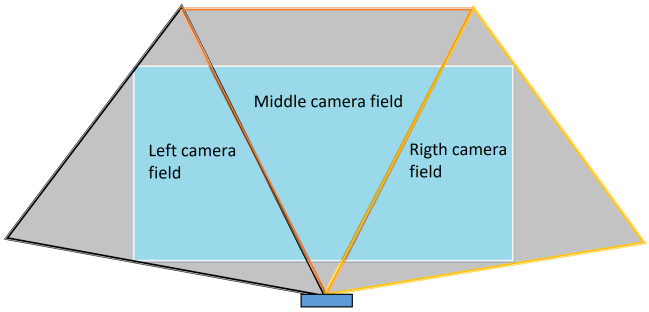
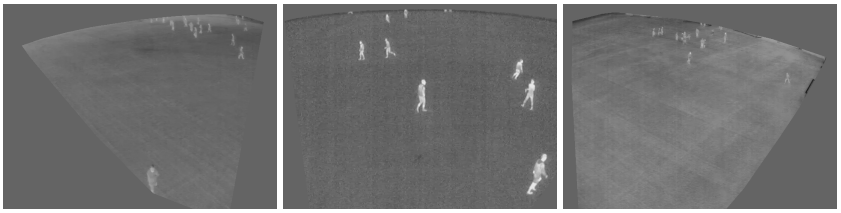



**Fig. 9.1:** Camera coverage for NFB-8, NFB-11, Aalborg Chang, Gug Boldklub, IAF and SSB. Blue represents the coverage area, and grey represents the segmented area.



**(a)** Left camera image **(b)** Middle camera image **(c)** Right camera image

**Fig. 9.2:** Example images from NFB-8, NFB-11, Aalborg Chang, Gug Boldklub, IAF and SSB.

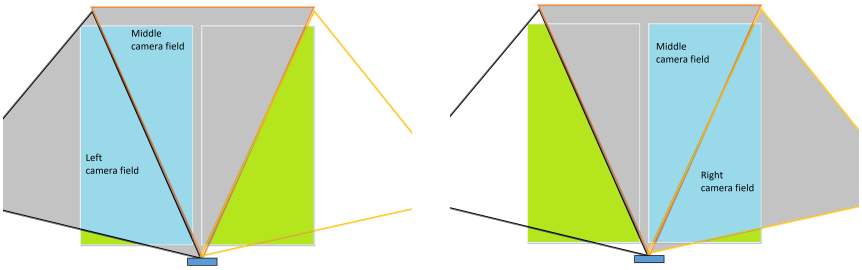
# 3 Algorithm for person detection in the fields

The algorithm from chapter 6 is used for person detection in thermal images, i.e. YOLO v3-based deep neural network. The network used two-step transfer learning. One is from RGB labelled data, and the second is from perfect thermal indoor soccer data. Thoroughly studied and carefully selected data with labels from overall recordings is then added to the network to learn and detect a person in outdoor soccer fields. An example is shown in Fig. 9.7.

# 4 Evaluation

The detection algorithm is evaluated on 1000 images, randomly selected from all data recordings. All the images are manually labelled and checked twice for ambiguities. Finally, the number of players in the field found by the algorithm is compared with the number of players counted manually. An example result is shown in Fig. 9.8.

Sensitivity, Equ. 9.1, is chosen as a performance evaluation measure as we are interested in positive detection. The desired algorithm, on average,




**(a)** Left and middle camera coverage for AAB-1 **(b)** Right and middle camera coverage for AAB-2

**Fig. 9.3:** Camera coverage for AAB-1 and AAB-2. Blue represents the coverage area, and grey represents the segmented area. Green is the covered area from other soccer fields being segmented out

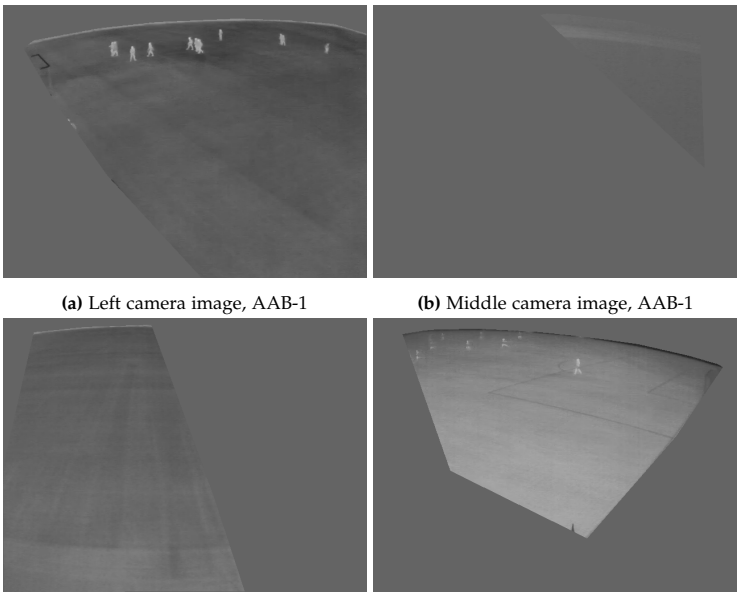
detects 73% of the players correctly. This implies that there is a 0.98% chance of getting the wrong detection for one image.

$$Sensitivity = \frac{TP}{TP + FP} \tag{9.1}$$

It is noticed that the quality of the image significantly decreases the detection rate (Fig. 9.9). Most of the time, the bad results are from blurred images, especially for players in the far regions. The blurred vision in the camera appears at the start of every video sequence. The cameras used in this setup record five minutes of sequence. So, blurred images are captured at the beginning of every new sequence.

## 5 Final results to the Municipality

The results are presented to the Municipality in tabular form with a colour representation. The results are calculated for each quarter that consists of 100 sequential images. For the fields covered with more than one camera, the number of players is added up and averaged up for 100 images. The colours in the table represent the level of occupancy in the fields ( Fig. **??**). The rows represent the time in hours duration from 7:00 am to 11:00 pm. Columns are the days of the week that are further divided into registered bookings vs the actual show-up at the fields. The report is submitted in the Danish language, and an example is shown in Fig. 9.11.



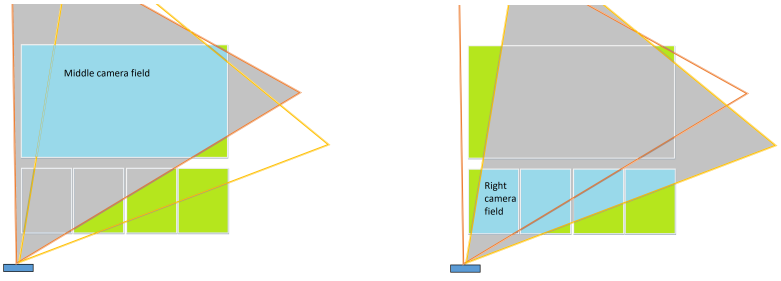
**(a)** Left camera image, AAB-1

**(b)** Middle camera image, AAB-1

**(c)** Middle camera image, AAB-2

**(d)** Right camera image, AAB-2

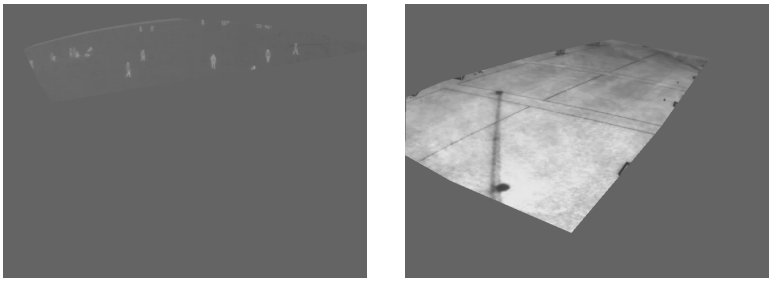
**Fig. 9.4:** Example images from AAB-1 and AAB-2.




**(a)** Middle camera coverage for SGI-1

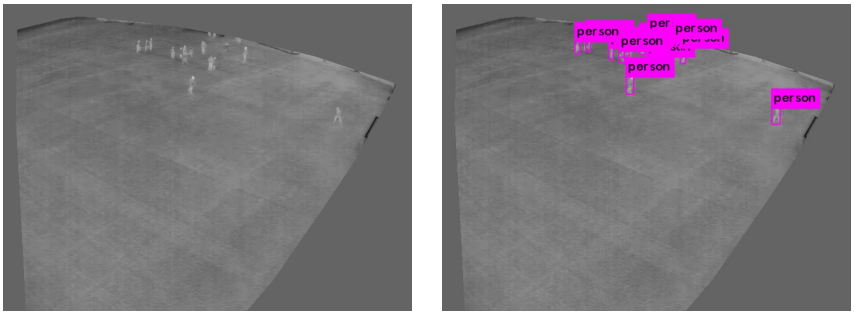
**(b)** Right camera coverage for SGI-2

**Fig. 9.5:** Camera coverage for SGI-1 and SGI-2. Blue represents the coverage areas, and grey represents the segmented area. Green is the covered area from other soccer fields being segmented out.



**(a)** Middle camera image **(b)** Right camera image

**Fig. 9.6:** Example images from SGI-1 and SGI-2 respectively.




**(a)** Example image from recording. **(b)** Positive person detections in the image.

**Fig. 9.7:** Example of person detection using the algorithm from chapter 6.

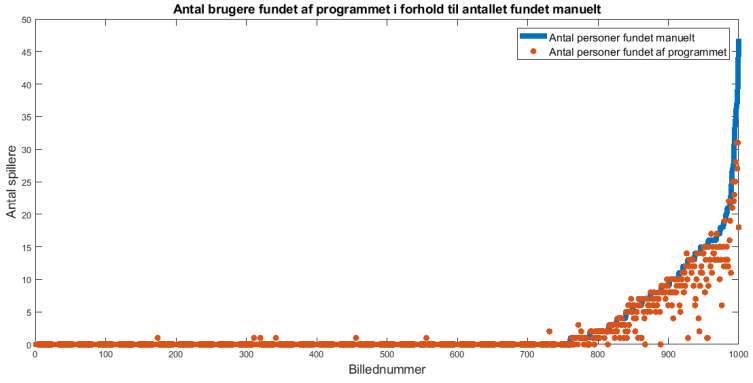
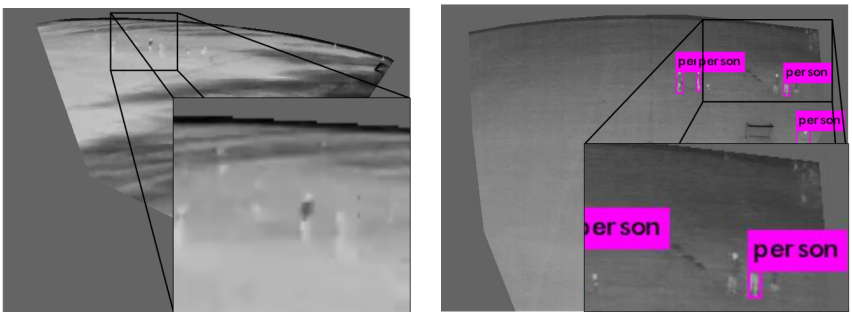



**Fig. 9.8:** The plot shows the number of players detected by the program vs the number of players labelled manually.




**(a)** Example of low-quality image and detection result

**(b)** Example of multiple body temperatures in one image.

**Fig. 9.9:** Example of blurred images and wrong detections.

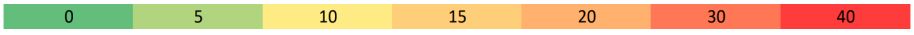



**Fig. 9.10:** Color representation for occupancy in the field. The number on the colours is the number of persons

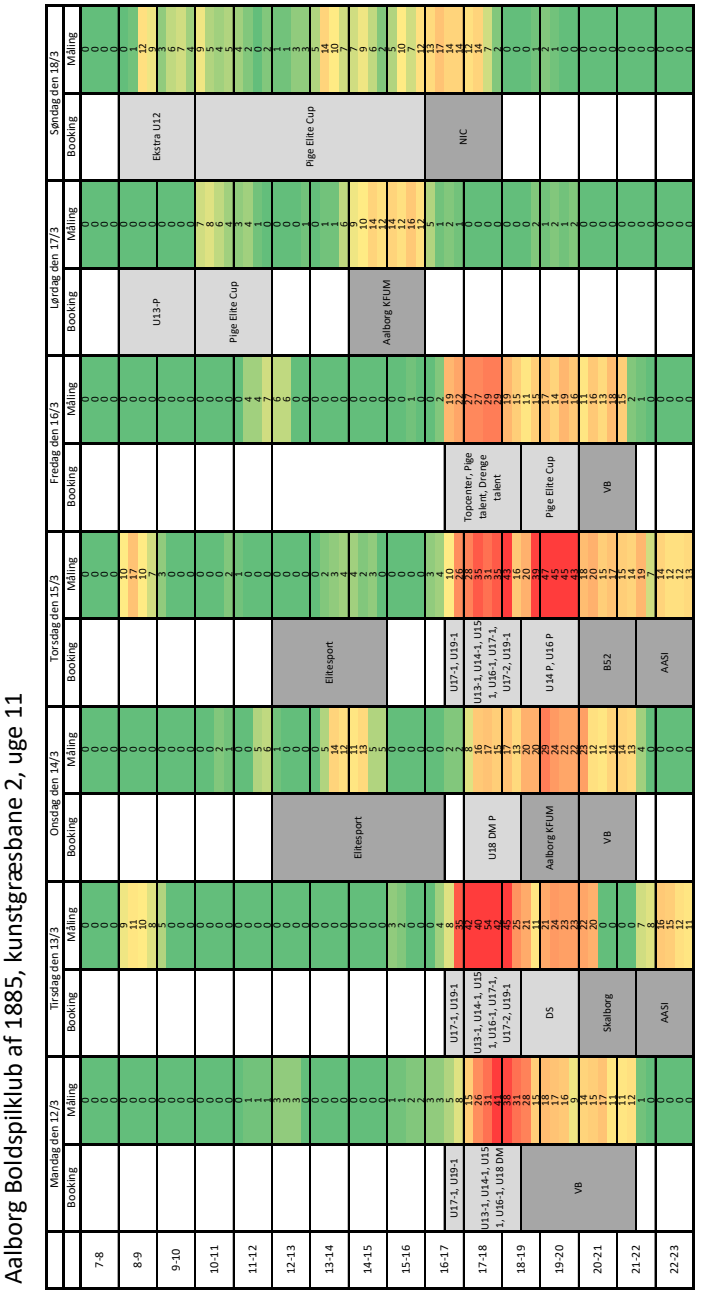



**Fig. 9.11:** Example Report.

# Chapter 10

# Conclusion and Discussion

## 1 Conclusion

This thesis covers the possible solutions for occupancy analysis in outdoor soccer fields. To that end, different camera setups are investigated for the application of person/player detection. Furthermore, this work presented challenges associated with camera setups like occlusion and distortion and outdoor monitoring like weather, environment, and light conditions.

With the focus on occupancy analysis, the first part of the thesis is a preliminary study that is aimed to observe and analyze the behaviour of five camera setups for player detection and counting. Each camera setup was closely studied for its installation complexity, cost-effectiveness, setup resilience for capability for long-term recordings and their behaviour in the wind, shadow, day and night recordings. In addition, a person's appearance and artefacts related to light and weather are studied for each camera output.

The camera setup analysis showed that the thermal camera performs well in low light and night conditions, but the setup is more expensive and complex. Whereas wide-angle high-resolution camera setups are found to be cheap and less complicated, the setup does not ensure the privacy of the field users and hinders long-term recordings. One fisheye camera is both affordable and straightforward in installation. The setup also gives anonymity to the users due to low resolution, but the video data looks pretty distorted, especially at the corners. Night vision in fisheye at the corners of the field almost become invisible. One thermal panning solution performs well at night and is simple but does not provide full field coverage for the whole time. One fisheye and one thermal camera solution are neither expensive and not cheap, and it is suitable for both day and night video feeds and preserves privacy. The problem with this kind of setup is the installation and data processing complexity. Any further comment on the setup required testing

for performance analysis. Three out of five camera setups were chosen for further research by analyzing the requirements for outdoor recording for a long time and considering privacy preservation. The following parts of this section address the challenges related to each chosen camera setup for player detection and occupancy analysis.

In the second part of the thesis, the results of one fisheye and wide-angle lens camera are presented for player detection and counting for occupancy analysis. The algorithm is based on adaptive background subtraction for detecting region of interest proposals as players. Afterwards, players are classified by using appearance-based features. It is observed that despite the high distortion along the corner of the field, the proposed method for occupancy analysis manages to achieve good performance for a reasonable period. The algorithm is tested for consecutive frames with almost the same light conditions. Results may vary by changing the light and environmental conditions.

The later part of the thesis presents player detection methods using a thermal camera setup. The first study proposes a machine learning approach to classify occluded vs non-occluded players. In the occluded blob, the number of players, i.e. 1, 2, 3, or 4, was identified by designing a virtual players and field setup. The method performed well for occupancy analysis in good thermal data. In the second study, many available thermal datasets are reviewed to use a deep neural network, and a lack of a diverse dataset is reported. So, thermal data of 20 weeks is thoroughly studied and categorized. Significant variance is observed in long-term recording, raising many more research questions. A diverse thermal dataset for person detection was finally presented. Moreover, the effect of each category of data was observed for training a deep learning network. The third study studied the influence of data harmonization on detection performance using a deep neural network. It is presented that the data homogenization-based preprocessing method does not improve the performance in person detection.

For the last study of camera setups, a combination of thermal and fisheye cameras is employed for detecting and counting players on a soccer field. First, a teacher-student-based network is employed to learn the thermal representations for the missing fisheye view. Both camera information was then utilized for detection using the deep network. It is observed that the student network starts learning and even sometimes performs better as it trains for more time. This leads to the conclusion that by increasing the amount of data, results get better while training online.

The final report to the municipality is then presented for long-term occupancy analysis of soccer fields using three thermal camera setups. The camera setup is mainly chosen due to its simplicity in installation and Independence on the artificial light conditions in the soccer fields. Overall each camera setup has its advantages and disadvantages. Fisheye camera

setup [1] performed well in daylight. Thermal camera setup performed well in low light conditions [2] and even in diverse environmental conditions [3]. The systems are simple to install and require minimum time sync for data processing for further analysis. The combination setup [4] can perform well in both day and night conditions provided enough online data streaming but it requires complex data sync, not complying with this condition can lead to significant performance degradation.

# 2 Outlook limitations and future perspectives

This thesis explores different methods and camera setups for occupancy analysis of outdoor soccer fields. Diverse outdoor challenges in connection to the outdoor environment are addressed in all sections of the thesis. Studying the abilities and limitations of each camera setup has opened the doors for many more possible research directions, not limited to soccer analysis. The investigation can be extended to functional application areas such as pedestrian analysis, security and surveillance, crowd analysis and many more.

The initial study for camera setup selection covers different outdoor parameters to be considered in accordance with the application area before the final setup. In the particular case of occupancy analysis, coverage area, setup simplicity, camera visibility, and long-time recording robustness are the most critical parameters. The study provides a guideline for the steps that could be considered before installing an outdoor setup for practical long-term recording applications. The study's limitation lies in the recording time, as the initial research is conducted for one day only. Therefore, it does not cover many more outdoor parameters encountered in long-term recording, e.g. camera behaviours for snow and high environmental temperatures. Furthermore, the recordings were made in moderate temperatures during the autumn season. Every season could have different environmental parameters to be considered before planning to have the long-term recording for that particular season.

Working with the fisheye camera has a clear advantage for the coverage of the whole field. In addition, using this camera reduces the installation and processing complexity as well as the cost. The algorithm for occupancy analysis using a fisheye camera yielded considerable results for controlled light conditions. The limitation of the fisheye lens lies in distortion around the image corners. No matter the algorithm, fisheye lens image suffers from severe distortion that, if combined with low light conditions, can lead to total blindness at the corners of the fields. For future work, better methods for distortion correction and image enhancement techniques can be explored for person detection. With the availability of enough data, CNN-based methods can also be inspected for better performance.

Thermal camera has the advantage of working well in low light condi-

tion and at night time. It also gives privacy preservation. The first work's limitation lies in the availability of enough outdoor thermal data. Every data set has its characteristics depending on the environment and the application area. No outdoor soccer field data is available online. That's why the studies generate artificial data by creating a virtual environment. The tests are performed on the data from one-day test recordings. Algorithm testing on long-term data can reveal more detailed results.

For the second study in the thermal domain, the experiments are conducted on long-term recordings, where the data is collected for 20 weeks. The long-term recordings lead to many new possible areas of research. It reveals many outdoor thermal-related issues that can not be encountered otherwise. The whole data is thoroughly studied and categorized. Categorizing the data enables the investigation of the effect of each kind of data on the results of detection. The results depict the contradicting behaviours while gathering opposite temperature images, i.e. images with a higher person's body temperature than the background and images with a lower person's body temperature than the background. Further studies and experimentation can reveal the actual effect of putting both kinds of images to gather and the effect of any prepossessing on the data. An intelligent selection of data for relevant applications may also lead to some exciting results to investigate through.

For the third study, the data homogenization-based preprocessing is investigated using the thermal person detection dataset. The study indicates that the polarity homogenization-based preprocessing does not improve the results. Therefore, other homogenization-based preprocessing techniques need to be tested to provide a general discussion on the matter.

Combining thermal and fisheye camera in a way that thermal camera covers only a part of the field is a novel and compelling study. The multi-modal and multi-view distillation cross-learn with time from thermal to fisheye. The principle of learning is based on a student-teacher-based network in which one framework acts as a student to learn from another framework we choose to act as a teacher. With time the student gets better and better at detection. The limitation of the work lies in the requirement of strict data synchronization, which is hard to achieve given that both cameras operate at different time stamps, frame rates and clip duration. More recordings with varying models of cameras may help to understand the options more clearly and make the data synchronization live and easy.

Every study has open up more research questions that may help improve the current state of the art. During the span of the PhD study, my research work revolves around particular camera model options. More camera model options for each setup can be investigated. The results may lead to either contradicting or confirming results. All the research conducted is based on self-collected data, and the outcome in the form of a report is also provided

to the administrations.

# References

[1] N. U. Huda, B. D. Hansen, R. Gade, and T. B. Moeslund, "Occupancy analysis of soccer fields using wide-angle lens," in *International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, 2017, pp. 354–359.

[2] N. U. Huda, K. H. Jensen, R. Gade, and T. B. Moeslund, "Estimating the number of soccer players using simulation-based occlusion handling," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2018, pp. 190 500–190 509.

[3] N. U. Huda, B. D. Hansen, R. Gade, and T. B. Moeslund, "The effect of a diverse dataset for transfer learning in thermal person detection," *Sensors*, vol. 20, no. 7, p. 1, Apr. 2020.

[4] A. Cioppa, A. Deliège, N. U. Huda, R. Gade, M. V. Droogenbroeck, and T. B. Moeslund, "Multimodal and multiview distillation for real-time player detection on a football field," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, June 2020.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-691-1

AALBORG UNIVERSITY PRESS