

UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR



PROYECTO FIN DE CARRERA

**DETECCIÓN Y SEGUIMIENTO EN VÍDEOS DEPORTIVOS
MULTICÁMARA**

Rafael Martín Nieto

Septiembre 2012

DETECCIÓN Y SEGUIMIENTO EN VÍDEOS DEPORTIVOS MULTICÁMARA

AUTOR: Rafael Martín Nieto
TUTOR: José María Martínez Sánchez



Video Processing and Understanding Lab
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Septiembre 2012

DETECTION AND TRACKING IN MULTI-CAMERA SPORTS VIDEOS

AUTHOR: Rafael Martín Nieto
ADVISOR: José María Martínez Sánchez



Video Processing and Understanding Lab
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
September 2012

Resumen

El objetivo principal de este proyecto es crear un sistema que, tras una sencilla configuración previa (y con cierta supervisión para los deportes de equipo), sea capaz de detectar y seguir cada jugador en la pista o campo de juego. Este sistema debe ser completo, general y modular, de forma que permita ser mejorado y modificado por trabajos futuros.

Después de analizar en detalle el estado del arte, se seleccionan algoritmos y módulos disponibles que sean útiles para el proyecto y se procede a su integración para el sistema final.

El sistema comenzará con un sistema base originalmente diseñado para video vigilancia, al que se le añadirán nuevas características y mejoras. Adicionalmente, se crearán los nuevos módulos necesarios para el correcto funcionamiento de todo el sistema.

Los deportes para los que se pretende desarrollar el sistema son los deportes individuales (por ejemplo, tenis) y deportes de equipo (por ejemplo, baloncesto y fútbol). Dadas las características muy diferenciadas de cada tipo de deporte, se desarrollará un sistema diferente para cada uno de ellos.

Para deportes individuales se desarrolla un sistema en el que todas las cámaras graban a un único jugador. Utilizando esta ventaja para realizar la fusión entre las distintas cámaras se facilita el proceso.

Para deportes de equipo, además del sistema de detección y seguimiento se diseña un sistema de evaluación de forma que se puedan obtener resultados cuantitativos del funcionamiento del sistema.

Palabras clave

Detección de objetos, seguimiento de objetos, sistemas multicámara, fusión, homografía, vídeos deportivos.

Abstract

Main objective of this project is to create a system which after a simple previous configuration (and with some supervision for team sports) is able to detect and track each player on the court or field. This system should be complete, general and modular, to be improved and modified by future work.

After analyzing in detail the state of the art, available algorithms and modules that are useful for the project are selected and will be integrated to the final system.

System starts with a base system, originally designed for video surveillance, which will be added new features and improvements. Additionally, new necessary modules for the proper functioning of the full system will be created.

Target sports to develop the system are individual sports (e.g., tennis) and team sports (e.g., basketball and football). Given the very different characteristics of each type of sport, a different system for each of them will be developed.

For individual sports, a system in which all the cameras record to a single player is developed. Using this advantage facilitates the fusion process.

For team sports, in addition to the detection and tracking system, an evaluation system is designed which allows to obtain quantitative results of the system performance.

Key words

Object detection, object tracking, multi-camera systems, fusion, homography, sports videos.

Agradecimientos

En primer lugar a Chema por su interés y ayuda en mi formación, no solo durante este proyecto si no desde el primer año de carrera. Gracias a él mi formación ha sido más completa y gratificante que la que se consigue únicamente en la carrera.

También a Juan Carlos por sus constantes consejos, formación y orientación, estando dispuesto a echar una mano cuando hace falta.

Agradecer también a mis padres, hermana y familia, por estar siempre a mi lado y apoyarme. Las comidas familiares de los fines de semana son momentos únicos que nunca olvidaré.

A Eva, por complementarme. Gracias a tu cariño has conseguido que estos años se pasen sin darme cuenta.

También mencionar a mis compañeros, en especial a Herrero, por haber hecho más ameno el día a día durante estos cinco años.

Finalmente a mis amigos. Una persona no es completa sin unos buenos amigos que te acompañen a lo largo del camino.

Rafael Martín
Septiembre de 2012

INDEX

1 Introduction	1
1.1 Motivation	1
1.2 Objectives	1
1.3 Structure of the document.....	2
2 State of the art.....	5
2.1 Introduction to sports video analysis	5
2.2 Sport video analysis techniques.....	7
2.2.1 Background extraction.....	7
2.2.2 Foreground analysis and blob detection	9
2.2.3 Team identification.....	10
2.2.4 Player tracking	11
2.2.5 Ball tracking	13
2.2.6 Trajectory association and fusion	15
2.2.6.1 Appearance-based approaches.....	17
2.2.6.2 Geometry-based approaches.....	17
2.2.6.3 Hybrid approaches.....	17
2.2.7 Overview of techniques	18
Table 2-1: Overview of sport video analysis techniques.....	19
2.3 Content sets	20
2.3.1 ISSIA Soccer Dataset	20
2.3.2 TRICTRAC Test Case Scenarios	21
2.3.3 PETS Football Dataset.....	22
2.3.4 APIDIS Basketball Dataset	24
2.3.5 3DLife ACM Multimedia Grand Challenge 2010 Dataset.....	26
2.3.6 CVBASE '06 Dataset.....	27
2.3.7 Overview of content sets	29
2.4 Commercial products.....	30
2.4.1 TRACAB Image Tracking System.....	30
2.4.2 AmiscoPro	31
2.4.3 Vis Track	31
3 Base system and common modules	33
3.1 Introduction	33
3.2 Base system overview	33
3.3 Base system modifications	34
3.3.1 Generation of background	34
3.3.1.1 Junction of portion from different frames	35
3.3.1.2 Mode of pixels from a set of frames.....	36
3.3.2 Extraction of the base mid-point	38

3.3.3	Generation of the file with coordinates	38
3.3.4	Adjustment of parameters to the characteristics of the videos	39
3.3.5	Addition of a mask for limiting the tracking area	41
3.4	Homographies.....	41
3.5	Field representation	43
4	Individual sports	45
4.1	Introduction	45
4.2	System	45
4.3	Fusion	46
4.4	Adjustments, testing and results	47
4.5	Discussion.....	50
5	Team sports	51
5.1	Introduction	51
5.2	System	51
5.3	Fusion	52
5.3.1	Basic fusion	54
5.3.2	Fusion using color information.....	54
5.3.3	Fusion adjusting correspondence between homographies.....	54
5.4	Evaluation system.....	55
5.5	Adjustments, testing and results	57
5.5.1	Fusion incremental development.....	58
5.5.2	Steps followed for each validation scenario	60
5.5.3	GT tracking based validation scenario	62
5.5.4	Base system tracking based validation scenario.....	65
5.5.5	GT tracking results vs. base system tracking results.....	68
6	Statistics from system results	69
7	Conclusions and Future Work	73
7.1	Conclusions	73
7.2	Future Work.....	74
	References	77
Annexes I		
A.	Base System.....	I
B.	2D Projective transformations: homographies	VII
C.	Generated backgrounds	IX
D.	Generated football masks	XIII
E.	Representation of blobs belonging to the same player after obtaining automatically the unique ID of the base system blobs	XVII

F.	Results of team sports using ground truth tracking	XXIII
G.	Results of team sports using base system tracking.....	XXVII
H.	Results of team sports. Contrast between ground truth tracking and base system tracking	XXXI
I.	Player statistics extracted from the results of the systems.....	XXXV
J.	Resulting trajectories for the football system	XLIX
K.	Supervised association of the trajectory fragments of each player for the football system	LIII
L.	Introducción.....	LXVII
M.	Conclusiones.....	LXXI
N.	Publicaciones generadas	LXXIII

FIGURE INDEX

FIGURE 2-1: BLOCK DIAGRAM OF THE CANONICAL SYSTEM.....	7
FIGURE 2-2: PLAYFIELD DETECTION RESULT: ORIGINAL FRAME (LEFT) AND THE DETECTED PLAYFIELD PIXELS (RIGHT).....	8
FIGURE 2-3: EXAMPLE OF VERTICAL INTENSITY DISTRIBUTION	10
FIGURE 2-4: PLAYER DETECTION AND TRACKING FROM A SINGLE CAMERA	12
FIGURE 2-5: SOME TYPICAL BALL SAMPLES IN BROADCAST SOCCER VIDEO (EXTRACTED FROM [8])	13
FIGURE 2-6: SAMPLE PROJECTION OF DETECTION MASK FROM MULTIPLE CAMERAS TO A TOP VIEW	15
FIGURE 2-7: POSITIONS OF THE SIX CAMERAS OF ISSIA SOCCER DATASET.....	21
FIGURE 2-8: EXAMPLES OF THE DIFFERENT VIDEOS OF ISSIA SOCCER DATASET.....	21
FIGURE 2-9: EXAMPLES OF DIFFERENT VIDEOS OF TRICTRAC TEST CASE SCENARIOS	22
FIGURE 2-10: POSITIONS OF THE CAMERAS OF PETS FOOTBALL DATASET	23
FIGURE 2-11: EXAMPLES OF THE DIFFERENT VIDEOS OF PETS FOOTBALL DATASET. TRAINING UP AND TESTING DOWN.....	23
FIGURE 2-12: EXAMPLES OF THE DIFFERENT VIDEOS OF APIDIS BASKETBALL DATASET	25
FIGURE 2-13: POSITIONS OF THE CAMERAS OF APIDIS BASKETBALL DATASET	26
FIGURE 2-14: POSITIONS AND EXAMPLES OF THE DIFFERENT VIDEOS OF 3DLIFE DATASET.....	27
FIGURE 2-15: EXAMPLES OF THE DIFFERENT VIDEOS OF CVBASE BASKETBALL DATASET	28
FIGURE 2-16: EXAMPLES OF THE DIFFERENT VIDEOS OF CVBASE SQUASH DATASET	28
FIGURE 2-17: EXAMPLES OF THE DIFFERENT VIDEOS OF CVBASE HANDBALL DATASET.....	29
FIGURE 3-1: BLOCK DIAGRAM OF THE BASE SYSTEM[24].....	33
FIGURE 3-2: EXAMPLE OF INCORRECT BACKGROUND.....	35
FIGURE 3-3: EXAMPLE OF CORRECT BACKGROUND	36
FIGURE 3-4: EXAMPLE (DETAIL) OF MALFUNCTION IN BACKGROUND GENERATOR	37
FIGURE 3-5: EXAMPLE OF INCORRECT BACKGROUND.....	37
FIGURE 3-6: EXAMPLE OF CORRECT BACKGROUND	37

FIGURE 3-7: EXAMPLES OF BOUNDING BOX (YELLOW AND RED) INDICATING THE BASE MID POINT (B) AND POINTS P1 AND P2.....	38
FIGURE 3-8: EXAMPLE OF SOME LINES OF THE GENERATED FILE IN TRACKING.....	39
FIGURE 3-9: EXAMPLE OF HOMOGRAPHY: (A) IMAGE PLANE, (B) TOP VIEW	42
FIGURE 3-10: FOOTBALL FIELD	44
FIGURE 3-11: BASKETBALL FIELD.....	44
FIGURE 3-12: TENNIS FIELD	44
FIGURE 4-1: BLOCK DIAGRAM OF THE SYSTEM FOR INDIVIDUAL SPORTS.....	45
FIGURE 4-2: IMAGE OF THE RESULTING TRAJECTORIES FROM EACH CAMERA.....	48
FIGURE 4-3: IMAGE OF THE RESULTING TRAJECTORY FROM FUSION	48
FIGURE 4-4: EXAMPLES OF THE POSSIBLE NUMBER OF CAMERAS TRACKING SIMULTANEOUSLY THE PLAYER: (A) ONE CAMERA, (B) TWO CAMERAS, (C) THREE CAMERAS.....	49
FIGURE 4-5: EXAMPLES OF THE MAJOR OR MINOR ERROR DEPENDING ON THE POSITION OF THE PLAYER	49
FIGURE 5-1: SYSTEM BLOCK DIAGRAM FOR TEAM SPORTS	51
FIGURE 5-2: EXAMPLE OF LOA.....	53
FIGURE 5-3: EXAMPLE OF TRAJECTORIES BEFORE APPLYING THE HOMOGRAPHY WITHOUT CORRECTION	59
FIGURE 5-4: EXAMPLE OF TRAJECTORIES AFTER APPLYING THE HOMOGRAPHY WITH CORRECTION	60
FIGURE 5-5: TOP VIEW TRACKING FOR EACH CAMERA	62
FIGURE 5-6: PRECISION AND RECALL FROM FUSION OF FACING CAMERAS WITH GROUND TRUTH TRACKING	63
FIGURE 5-7: RESULTING TRACKINGS FROM THE FUSION OF FACING CAMERAS	64
FIGURE 5-8: PRECISION AND RECALL FROM FUSION OF THE DIFFERENT RESULTING REGIONS WITH GROUND TRUTH TRACKING.....	64
FIGURE 5-9: RESULTING TRACKING FROM THE REGION FUSION	65
FIGURE 5-10: PRECISION AND RECALL FROM FUSION OF FACING CAMERAS WITH BASE SYSTEM TRACKING	66
FIGURE 5-11: PRECISION AND RECALL FROM FUSION OF THE DIFFERENT RESULTING REGIONS WITH BASE SYSTEM TRACKING	67

FIGURE 6-1: EXAMPLES OF THE ZIGZAG EFFECT IN THE TRAJECTORIES OF THE TENNIS SYSTEM (A) AND IN THE TRAJECTORIES OF THE FOOTBALL SYSTEM (B)	69
FIGURE 6-2: EXAMPLE OF THE RESULTING TENNIS STATISTICS VIDEO	70
FIGURE 6-3: EXAMPLE OF THE RESULTING FOOTBALL STATISTICS VIDEO	70
FIGURE A-1: BLOCK DIAGRAM OF THE FOREGROUND SEGMENTATION MODULE	II
FIGURE A-2: FOREGROUND MASK AT DIFFERENT STAGES: INITIAL FOREGROUND MASK (A), AFTER SHADOW REMOVAL (B) AND AFTER NOISE REMOVAL (C)	III
FIGURE B-1: EUCLIDEAN PROJECTIVE TRANSFORMATION	VIII
FIGURE C-1: BACKGROUND OF CAMERA 1	IX
FIGURE C-2: BACKGROUND OF CAMERA 3	IX
FIGURE C-3: BACKGROUND OF CAMERA 4	IX
FIGURE C-4: BACKGROUND OF CAMERA 5	IX
FIGURE C-5: BACKGROUND OF CAMERA 6	X
FIGURE C-6: BACKGROUND OF CAMERA 8	X
FIGURE C-7: BACKGROUND OF CAMERA 1	X
FIGURE C-8: BACKGROUND OF CAMERA 2	XI
FIGURE C-9: BACKGROUND OF CAMERA 3	XI
FIGURE C-10: BACKGROUND OF CAMERA 4	XI
FIGURE C-11: BACKGROUND OF CAMERA 5	XII
FIGURE C-12: BACKGROUND OF CAMERA 6	XII
FIGURE D-1: MASK FOR CAMERA 1	XIII
FIGURE D-2: MASK FOR CAMERA 2	XIII
FIGURE D-3: MASK FOR CAMERA 3	XIV
FIGURE D-4: MASK FOR CAMERA 4	XIV
FIGURE D-5: MASK FOR CAMERA 5	XV
FIGURE D-6: MASK FOR CAMERA 6	XV
FIGURE E-1: TRAJECTORY FOR UNIQUEID = 200	XVII

FIGURE E-2: TRAJECTORY FOR UNIQUEID = 201	XVII
FIGURE E-3: TRAJECTORY FOR UNIQUEID = 202	XVIII
FIGURE E-4: TRAJECTORY FOR UNIQUEID = 1	XVIII
FIGURE E-5: TRAJECTORY FOR UNIQUEID = 105	XVIII
FIGURE E-6: TRAJECTORY FOR UNIQUEID = 5	XVIII
FIGURE E-7: TRAJECTORY FOR UNIQUEID = 6	XVIII
FIGURE E-8: TRAJECTORY FOR UNIQUEID = 8	XVIII
FIGURE E-9: TRAJECTORY FOR UNIQUEID = 9	XIX
FIGURE E-10: TRAJECTORY FOR UNIQUEID = 10	XIX
FIGURE E-11: TRAJECTORY FOR UNIQUEID = 11	XIX
FIGURE E-12: TRAJECTORY FOR UNIQUEID = 14	XIX
FIGURE E-13: TRAJECTORY FOR UNIQUEID = 20	XIX
FIGURE E-14: TRAJECTORY FOR UNIQUEID = 21	XIX
FIGURE E-15: TRAJECTORY FOR UNIQUEID = 26	XX
FIGURE E-16: TRAJECTORY FOR UNIQUEID = 104	XX
FIGURE E-17: TRAJECTORY FOR UNIQUEID = 107	XX
FIGURE E-18: TRAJECTORY FOR UNIQUEID = 108	XX
FIGURE E-19: TRAJECTORY FOR UNIQUEID = 113	XX
FIGURE E-20: TRAJECTORY FOR UNIQUEID = 114	XX
FIGURE E-21: TRAJECTORY FOR UNIQUEID = 120	XXI
FIGURE E-22: TRAJECTORY FOR UNIQUEID = 125	XXI
FIGURE E-23: TRAJECTORY FOR UNIQUEID = 126	XXI
FIGURE E-24: TRAJECTORY FOR UNIQUEID = 128	XXI
FIGURE E-25: TRAJECTORY FOR UNIQUEID = 155	XXI
FIGURE F-1: PRECISION AND RECALL FROM FUSION OF CAMERAS 1 AND 2	XXIV
FIGURE F-2: PRECISION AND RECALL FROM FUSION OF CAMERAS 3 AND 4	XXIV

FIGURE F-3: PRECISION AND RECALL FROM FUSION OF CAMERAS 5 AND 6	XXV
FIGURE F-4: PRECISION AND RECALL FROM FUSION OF FACING CAMERAS	XXV
FIGURE F-5: PRECISION AND RECALL FROM FUSION OF THE DIFFERENT RESULTING REGIONS .	XXVI
FIGURE G-1: PRECISION AND RECALL FROM FUSION OF CAMERAS 1 AND 2.....	XXVIII
FIGURE G-2: PRECISION AND RECALL FROM FUSION OF CAMERAS 3 AND 4.....	XXVIII
FIGURE G-3: PRECISION AND RECALL FROM FUSION OF CAMERAS 5 AND 6.....	XXIX
FIGURE G-4: PRECISION AND RECALL FROM FUSION OF FACING CAMERAS	XXIX
FIGURE G-5: PRECISION AND RECALL FROM FUSION OF THE DIFFERENT RESULTING REGIONS ..	XXX
FIGURE H-1: PRECISION AND RECALL FROM FUSION OF CAMERAS 1 AND 2.....	XXXII
FIGURE H-2: PRECISION AND RECALL FROM FUSION OF CAMERAS 3 AND 4.....	XXXII
FIGURE H-3: PRECISION AND RECALL FROM FUSION OF CAMERAS 5 AND 6.....	XXXIII
FIGURE H-4: PRECISION AND RECALL FROM FUSION OF FACING CAMERAS	XXXIII
FIGURE H-5: PRECISION AND RECALL FROM FUSION OF THE DIFFERENT RESULTING REGIONS	XXXIV
FIGURE I-1: STATISTICS FOR THE TENNIS PLAYER	XXXV
FIGURE I-2: STATISTICS FOR UNIQUEID = 200	XXXVI
FIGURE I-3: STATISTICS FOR UNIQUEID = 201	XXXVII
FIGURE I-4: STATISTICS FOR UNIQUEID = 202	XXXVII
FIGURE I-5: STATISTICS FOR UNIQUEID = 1	XXXVIII
FIGURE I-6: STATISTICS FOR UNIQUEID = 105	XXXVIII
FIGURE I-7: STATISTICS FOR UNIQUEID = 5	XXXIX
FIGURE I-8: STATISTICS FOR UNIQUEID = 6	XXXIX
FIGURE I-9: STATISTICS FOR UNIQUEID = 8	XL
FIGURE I-10: STATISTICS FOR UNIQUEID = 9	XL
FIGURE I-11: STATISTICS FOR UNIQUEID = 10	XLI
FIGURE I-12: STATISTICS FOR UNIQUEID = 11	XLI
FIGURE I-13: STATISTICS FOR UNIQUEID = 14	XLII

FIGURE I-14: STATISTICS FOR UNIQUEID = 20	XLII
FIGURE I-15: STATISTICS FOR UNIQUEID = 21	XLIII
FIGURE I-16: STATISTICS FOR UNIQUEID = 26	XLIII
FIGURE I-17: STATISTICS FOR UNIQUEID = 104	XLIV
FIGURE I-18: STATISTICS FOR UNIQUEID = 107	XLIV
FIGURE I-19: STATISTICS FOR UNIQUEID = 108	XLV
FIGURE I-20: STATISTICS FOR UNIQUEID = 113	XLV
FIGURE I-21: STATISTICS FOR UNIQUEID = 114	XLVI
FIGURE I-22: STATISTICS FOR UNIQUEID = 120	XLVI
FIGURE I-23: STATISTICS FOR UNIQUEID = 125	XLVII
FIGURE I-24: STATISTICS FOR UNIQUEID = 126	XLVII
FIGURE I-25: STATISTICS FOR UNIQUEID = 128	XLVIII
FIGURE I-26: STATISTICS FOR UNIQUEID = 155	XLVIII
FIGURE J-1: RESULTING TRAJECTORIES FOR THE IDEAL FUSION IN THE GROUND TRUTH TRACKING SCENARIO.....	L
FIGURE J-2: RESULTING TRAJECTORIES FOR THE EXPERIMENTAL FUSION IN THE GROUND TRUTH TRACKING SCENARIO	L
FIGURE J-3: RESULTING TRAJECTORIES FOR THE IDEAL FUSION IN THE BASE SYSTEM TRACKING SCENARIO.....	LI
FIGURE J-4: RESULTING TRAJECTORIES FOR THE EXPERIMENTAL FUSION IN THE BASE SYSTEM TRACKING SCENARIO	LI
FIGURE K-1: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 200	LIII
FIGURE K-2: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 201	LIV
FIGURE K-3: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 202	LIV
FIGURE K-4: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 1	LV
FIGURE K-5: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 105	LV

FIGURE K-6: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 5	LVI
FIGURE K-7: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 6	LVI
FIGURE K-8: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 8	LVII
FIGURE K-9: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 9	LVII
FIGURE K-10: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 10	LVIII
FIGURE K -11: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 11	LVIII
FIGURE K -12: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 14	LIX
FIGURE K-13: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 20	LIX
FIGURE K-14: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 21	LX
FIGURE K-15: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 26	LX
FIGURE K-16: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 104	LXI
FIGURE K-17: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 107	LXI
FIGURE K-18: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 108	LXII
FIGURE K-19: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 113	LXII
FIGURE K-20: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 114	LXIII
FIGURE K-21: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 120	LXIII
FIGURE K-22: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 125	LXIV
FIGURE K-23: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID = 126	LXIV

FIGURE K-24: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID
= 128..... LXV

FIGURE K-25: SUPERVISED ASSOCIATED FRAGMENTS AND COMPLETE TRAJECTORY FOR UNIQUEID
= 155..... LXV

TABLE INDEX

TABLE 2-1: OVERVIEW OF SPORT VIDEO ANALYSIS TECHNIQUES.....	19
TABLE 2-3: OVERVIEW OF CONTENT SETS	29
TABLE 5-1: NUMBER OF CORRECT FUSIONS.....	57
TABLE 5-2: NUMBER OF TRAJECTORY FRAGMENTS OF EACH PLAYER.....	68

1 Introduction

1.1 Motivation

The main motivation of this project is to realize a system able to detect and track the players on the field from videos taken from different positions via a multicamera system.

From the perspective of leisure, sport videos represent a significant proportion of the total broadcasts from public and commercial television. A lot of work has already been carried out in the analysis of sports video content, and work to improve and enrich the sports video is growing quickly because the high demand from users.

From a professional perspective, sports video analysis, especially in ball games (football, tennis, basketball, ...) is particularly useful for analyzing and improving tactics and players (and team) performance. With such system, coaches can obtain data, statistics and other information that are difficult to obtain without a system of this type. It is also interesting for sports broadcasts since it allows getting statistics and interesting facts for the audience.

1.2 Objectives

The main objective of this project is to create a system which, after a previous configuration (and with some supervision for team sports), is able to detect and track each player in the field. This system should be complete, general and modular, easing future work improvements and modifications.

The system will start with a base system originally designed for video surveillance, which will introduce new features and enhancements. Additionally, new modules needed for the proper operation of the entire system will be created.

The sports under consideration are both individual sports (e.g., tennis) and team sports (e.g., basketball, football). Given the mentioned video features, there will be a separate system for each of them.

After having obtained these basic systems, additional functionalities related to performance, statistics and different representation of the results are studied and developed.

1.3 Structure of the document

The memory includes the following chapters:

- **Chapter 1.** Motivation and objectives of the project.
- **Chapter 2.** State of the art of sport video analysis and introduction of some sport video content sets.
- **Chapter 3.** Description of the base system and the common modules, including base system modifications, homographies and scripts for the field representation.
- **Chapter 4.** Work developed for individual sports, describing the system, the fusion applied and the adjustments, testing and results.
- **Chapter 5.** Work developed for team sports, describing the system, the different fusions applied, the evaluation system designed and the adjustments, testing and results.
- **Chapter 6.** Statistics extracted from the results of the systems developed in chapters 4 and chapter 5.
- **Chapter 7.** Conclusions of the project and future work.
- **Annex A.** Base system described precisely and in detail.
- **Annex B.** Mathematical theory of homographies.
- **Annex C.** Generated backgrounds for each camera of the tennis and football datasets used.
- **Annex D.** Generated football masks.
- **Annex E.** Representation of blobs belonging to the same player after obtaining automatically the unique ID of the base system blobs.
- **Annex F.** Results of team sports using ground truth tracking.
- **Annex G.** Results of team sports using base system tracking.
- **Annex H.** Contrast between ground truth tracking results and base system tracking results in team sports.
- **Annex I.** Player statistics extracted from the results of the systems.

- **Annex J.** Figures of the resulting trajectories for the football system.
- **Annex K.** Supervised association of the trajectory fragments of each player for the football system

2 State of the art

2.1 Introduction to sports video analysis

Sports broadcasts constitute a major percentage in the total of public and commercial television broadcasts. A lot of work has already been carried out on content analysis of sports videos, and the work on enhancement and enrichment of sports video is growing quickly due to the great demands of customers.

In [1], an overview of sports video research is given, describing both basic algorithmic techniques and applications. Sports video research can be classified into the following two main goals: indexing and retrieval systems (based on high-level semantic queries) and augmented reality presentation (to present additional information and to provide new viewing experience to the users). The analysis of events can be carried out by combining various attributes of the video, including its structure, events and other content properties. In event detection, features can be extracted from three channels: video, audio and text. The definition of what is interesting differs for each viewer. While sport fans are more interested in events like goals or spectacular points in tennis, the coaches might be more interested in the errors of the players to help them improve their capabilities.

The different approaches and methods of designing the event detection algorithms are:

- low-level features vs. object-related features: low-level features are the cinematic features acquired directly from the input video, and object-related features are attributes of objects such as ball locations and player shapes, acquired by more complex algorithms.
- single channel vs. multiple channels: some events can be detected using the features from a single channel and for other events the features from multiple channels to be detected, usually video and audio, are needed.
- game-specific vs. generic: most of the developed event detectors are game-specific because different games have different events but obviously the objective is an algorithm which can detect the events of multiple games.
- feature model vs. shot pattern: some algorithms directly model the relations between the features and the events; but some events have no intuitive relations with the features requiring context information for the detection and use the shot patterns to capture the context information.

- learning-based vs. non-learning-based (named “normal” in [1]): learning-based algorithms capture the relations between the features and the events by statistical analysis and optimization. In general, the machine learning approach is chosen when the relations between the features and the events are difficult to define intuitively.
- broadcast video vs. non-broadcast video: most of the algorithms were designed for broadcast sports video from which they can use the editing information as an additional source of information.

In addition, algorithmic building blocks must consider structure analysis (for example of a match), object detection and segmentation (e.g., position of players, shape and movements of the athletes), ball and player tracking (specially the tracking of a small ball is a difficult problem), camera calibration (to determine the geometric mapping between the image coordinates and the real world positions and 3D reconstruction: multiple cameras can be used to obtain views from different position and compute a complete reconstruction of the playfield).

The main applications of sports video processing are:

- Abstracting: creating a shortened version of a video that still comprises the essential information from the complete video.
- Tactic and performance analysis: to understand the tactics that teams or individual players have used and to evaluate the performance of a team or a player through analyzing their motion and activity in games.
- Augmented reality presentation of sports: with two basic techniques: 3D reconstruction of the game to provide arbitrary views, and to insert some illustrations into the original video or to provide the illustration in extra windows to help consumers to understand the video easier.
- Sport video for small devices: the transmission of sport events on small devices like PDA, 3G/UMTS phones will become more popular in the future
- Referee assistance: to help the referee in cases of difficult decisions and partially replace the work of the referee.

There are two main types of sports video, used to perform the analysis: edited broadcast and multi-camera.

- Edited broadcast [2][3][4][5]: videos are edited for broadcasting on television. On this type of videos, there are scenes of the game, repetitions and changes of viewpoint of the observer. This type of videos is the most common of all.
- Multi-camera: In general, multi-camera videos are not edited. They are subdivided into two main classes:
 - Mobile [6]: Cameramen follow the players, changing the orientation of their cameras. The obtained videos usually are used to generate the broadcast videos, mixing and editing parts of them.
 - Fixed [7][8][9]: The cameras remain fixed from the beginning to the end of the recording. In this case, no operator is needed for controlling the orientation of the camera.

2.2 Sport video analysis techniques

The canonical system for detecting and tracking players in a field can be decomposed into several main blocks, as depicted in Figure 2-1. The different techniques and associated algorithms that are key for any sports video analysis system are described in the following subsections.

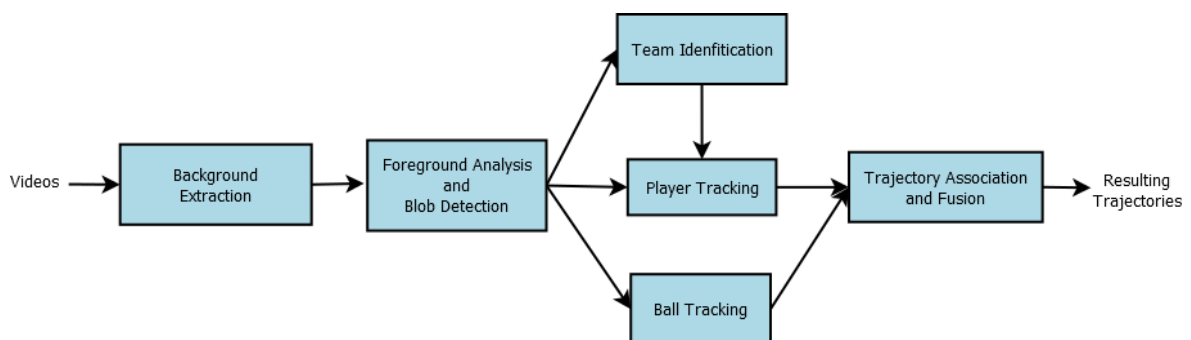


Figure 2-1: Block diagram of the canonical system

2.2.1 Background extraction

Background extraction is a simple and very common method used for moving objects segmentation which consists of the difference between a set of images and the background model. The background model is generated from an empty image of the field (if it is available) or from a fragment of the video with non-static players. Some of the main

problems in background extraction are the changes in the environment such as illumination, shadows or background moving objects.

Some statistical adaptive methods are used in [2] for this block. These methods update the background model considering the background pixel as a Gaussian distribution and work well in relatively simple scenarios.

A histogram learning technique is employed in [3] to detect the playfield pixels, using a training set of soccer videos. The pixels in the training set videos are labeled as playfield pixels and non-playfield pixels, either manually or using a semi-supervised method. Based on its RGB color vector, each labeled playfield pixel is placed into the appropriate RGB bin of the playfield histogram. A similar process is carried out for the pixels labeled as non-playfield. An example of the playfield detection result is shown in figure 2-2

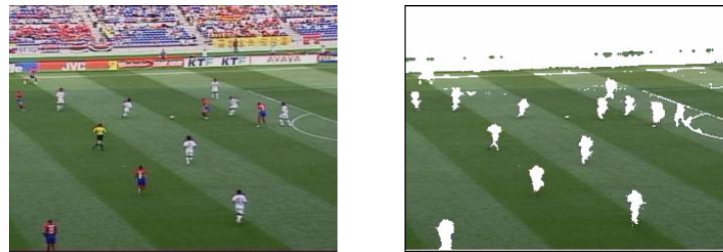


Figure 2-2: Playfield detection result: Original frame (left) and the detected playfield pixels (right)

In [5], background is modeled with a mixture of Gaussians and learned before a match without the need of an empty scene. Background detector is first used to extract a pitch mask for saving processing time and avoiding false alarms from the crowd. This pitch mask is computed using the hue histogram and the projection of the known pitch geometry. A background model is generated in [6], and a background updating procedure is used to continuously adapt the model to the variations in light condition.

In [7], the region of the ground field is extracted, first by segmenting out non-field regions like advertisements, and then the players are extracted on this field on the basis of the field extraction. The field has a uniform color of green and occupies large area in the image. So histograms of each color, R/G/B, are calculated and peak values are found from each. Using these peak values, the image is binarized, considering a threshold value for each color. After obtaining the binary image, morphological filtering is applied and the largest connected component is extracted, that represents the field but has many holes caused by colors of players. The field mask is obtained by filling the interior of this boundary.

The field is represented in [9] by a constant mean color value that is obtained through prior statistics over a large data set. The color distance between a pixel and the mean value of the field is used to determine whether it belongs to field or not. In the system, only hue and saturation components in HSV color space are computed to exclude effect of illumination.

2.2.2 Foreground analysis and blob detection

Blob detection is based on connecting the pixels of the foreground to identify the players. The goal is to adjust as much as possible to the contour of the player blob, separating the near or overlapping possible players. The correct determination of the number of components in a blob is important for making correct decisions during the tracking.

In [2], the nodes are grouped considering the edges between them. The area of the blobs is the parameter used to define the number of components of each node. The graph is constructed from the set of blobs obtained during the segmentation step in such a way that nodes represent blobs, and edges represent the distance between these blobs. In order to eliminate short connection, the blobs are considered as regions of the original images and a sequence of morphological operations is applied for post-processing and splitting these blobs. A blob in the current frame can be split in two or more blobs depending on the number of components of the corresponding node in the graph and on its configuration on the previous frame.

In [3], connected component analysis (CCA) scans the binary mask and groups its pixels into components based on pixel connectivity. Ideally, the non-playfield pixels inside the extracted playfield areas should be the foreground pixels that can be grouped into different foreground blobs by CCA. Since the ball is typically quite small (less than 30 pixels in size for QVGA content), erosion can not be used before grouping pixels into different blobs, but dilation can still be used. Therefore, noisy areas have to be removed later by shape analysis and true object appearance model.

The foreground mask is morphologically filtered in [4] (closing after opening) to get rid of outliers. This is followed by a connected component analysis that allows creating individual blobs and bounding boxes are created.

In [5], for an isolated player, the image measurement comes from the bottom of foreground region directly. The measurement covariance in an image plane is assumed to be a constant and diagonal matrix, because foreground detection in an image is a pixel-wise operation. Once some bounding edge of a target is decided to be observable, its opposite

unobservable bounding edge could be roughly estimated. Because the estimate is updated using partial measurements whenever available, it is more accurate and robust than using prediction only.

A further step with the connectivity analysis has been introduced in [6] to extract connected regions and remove small regions due to noise. This procedure scans the entire image and groups neighbor pixels into regions. The connectivity analysis eliminates shadows by using geometrical considerations about the shape of each region. Extensions of areas in the orthogonal directions with respect to the expected vertical position of players are removed during the construction of connected regions. At the end of this step each connected region, that has to be considered in the successive tracking procedure, is surrounded by a bounding box.

2.2.3 Team identification

The information of the uniform of the teams is used to identify the team of the blob and difference between players, goalkeepers or referees. Generally a player can be modeled as a group of many regions, each region having some predominant colors.

Dividing the model of the player in two or more regions is attempted in [2], so that each region represents a part of the team's uniform, e.g. t-shirt, short, socks. For each region, a filtering based on the vertical intensity distribution of the blobs is defined. Limit values are defined from the minimal and maximal mean value of the vertical distribution for each region. An example of vertical intensity distribution is shown in figure 2-3.

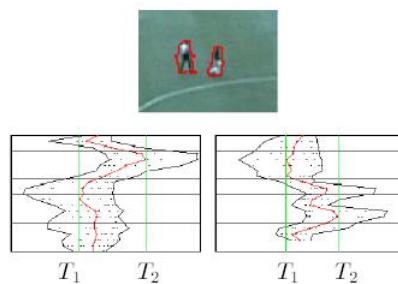


Figure 2-3: Example of vertical intensity distribution

Color histograms are used to classify the objects of templates for each class in [4]. The different classes consist of team 1, team 2, goalkeeper team 1, goalkeeper team 2, and referee. The center part of the detected blobs is taken and color histograms are calculated.

Histograms are compared, after normalization, with the color histograms of the templates using the Bhattacharyya distance. Automated clustering could be used, however this would increase the complexity of learning and produce more uncertain results. Selecting the templates can be very straightforward, (e.g., user input can be asked to classify the bounding boxes in the first frame).

This block can be also implemented using a histogram-intersection method[5]. The result for each player is a five-element vector, indicating the probability that the observed player is wearing one of the five categories of uniform (the categories are the same as described in [4]).

In [7], the group behavior of soccer players is analyzed using color histogram back-projection to isolate players on each team. But since different teams can have a similar histogram, vertical distribution of colors is used. After computing vertical distribution of RGB, it is compared with each team's model distribution. Model distribution is obtained when the first image is considered. The similarity measure is computed by convolution within a small range because the vertical location of a player in the template can vary slightly. A similar distribution is obtained when players of the same team are computed.

2.2.4 Player tracking

When the position of each player position is detected, the next objective is to track the player to know where the player is at each moment.

In [2], the tracking of each player is performed by searching an optimal path in the graph defined in blob detection. In order to start tracking a player, its initial blob is defined and the corresponding node is found on the graph. At each step, the minimal path in the graph is considered, using the distance information between the blobs. During an occlusion of two players, the direction of their trajectory is considered in order to decide the correct paths, specially, in cases when there are players in contact. The tracking of the players in case of contact or occlusions by other players is more difficult.

The bounding box and centroid coordinates of each player are used in [5] as state and measurement variables in a Kalman filter. For the multi-view tracking process, a three-step procedure is applied. The first step is to associate measurements to established tracks and update these tracks. The second step is to initialize tracks for the measurements unmatched to any existing track. Finally, the fixed population constraint for each category of players (ten outfield players and one goalkeeper per team, three referees) is used to recognize the

members in each category. In track update, each player is modeled as a track and has its state estimate updated, if possible, by an overall measurement fused for at least one camera. In track initialization, the measurements unmatched after checking them against existing tracks are checked against each other to find potential new tracks. Finally, in track selection, there is a procedure of tracking aided recognition for the 25 most likely players (11 from each team and 3 referees). Due to false alarms and tracking errors, there normally exist more than 25 tracks for the players. A player likelihood measure is calculated for each target on the basis of confidence of category estimate, number of support cameras, domain knowledge in positions (for goalkeepers and linesmen), frames of being tracked or missing, as well as the fixed population constraint. A fast sub-optimal search method gives reasonable results. An example of player detection and tracking is shown in figure 2-4.

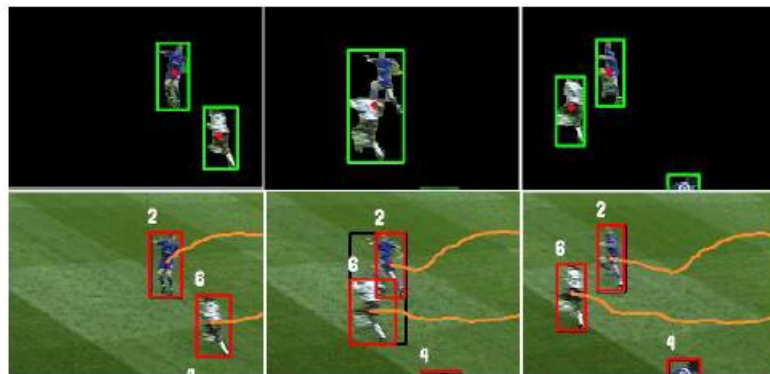


Figure 2-4: Player detection and tracking from a single camera

The multiplayer tracker may use the Maximum a Posteriori Probability[6] to get the best fit of state and observation sequences. The state vector includes information about the location, velocity, acceleration and dimension of the single bounding box. The configuration providing the Maximum a posteriori probability is the one selected as best fit between state observation and state prediction. At this point a further step for the prediction validation has to be carried out. By comparing the matches between the observations with the predictions, two possible situations may happen: there are some observations that do not match any prediction and there are some predictions that do not correspond to any observation. In the first case different situations may occur: 1) the observation could be a new entry blob if is on the image border, then a new blob is generated with an incoming state; 2) the observation could be a resumed blob if it is close to a prediction with a disappeared state; 3) the observation could be generated by noise, then a new entity is created and observed along a temporal window until a decision on its persistence is taken. In the second case, if the prediction has not a correspondent among

the observations and it is not on the image border (it is not in an outgoing situation), it means that the foreground segmentation step was not detecting the blob and then the status vector is maintained setting (disappeared blob). An additional analysis is required when merge event occurs between two or more blobs. Predict that two or more blobs will merge is possible, but since to maintain their vector status separated is needed, to logically split them in the corresponding observations. This splitting procedure could be difficult especially when two or more players are very close to each other and the occlusion is almost total.

Template matching and Kalman filter is used in [7]. The templates of players are extracted from player mask using connected component extraction. First, new players that do not significantly overlap with the bounding box of a player already tracked are found out. Then, the new players are inserted to tracking list. Location of players at the next frame is predicted by Kalman filter and template matching at that location is performed. Finally, the player template is updated. The main problem of player tracking is occlusion and in [7] only occlusion between different teams is considered.

In [10], starting from a frame where no two players occlude each other, players are tracked automatically by considering their 3D world coordinates. When a merging of two (or more) players is detected, separation is done by means of a correlation-based template matching algorithm.

2.2.5 Ball tracking

The method for ball tracking is similar to the method used for player tracking. But ball tracking is more difficult than player tracking because automatic detection of the ball is harder, due to its reduced dimensions in the image. If a player has the ball, tracking is difficult because the ball is frequently occluded by the player. Figure 2-5 shows some samples of the ball in broadcast soccer video, extracted from [8].

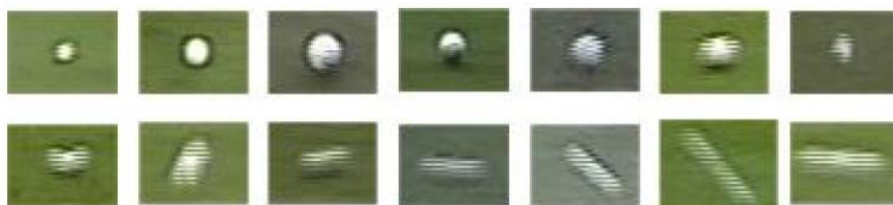


Figure 2-5: Some typical ball samples in broadcast soccer video (extracted from [8])

The differences and characteristic features of some systems are described below.

In [3], there is a predefined range for the ball blob's area according to the camera configuration. Meanwhile, the proportion of detected white pixels in the blob indicates the possibility of being a ball. Roundness and eccentricity for a blob candidate should be close to 1.0, different from disconnected segments of field lines.

Other option is, using a Hough transform over the foreground mask[4] to detect candidate pixels that correspond with circles.

Temporal information is used between the frames by creating a search window around the last found position of the ball. Only in this window, and only in the corresponding foreground blobs, the new ball location is searched. This allows preventing that the entire frame has to be searched for the ball location, avoiding costly processing. To overcome propagating errors due to a misdetection of the ball, every 10 frames the entire foreground frame is searched for the best ball candidate.

When the player is neighboring with the ball the detection drops. However this information can be exploited. The player which collides is denoted with the ball and the search window is extended to include the bounding box of the player. As such, the ball can be detected when the player passes the ball. Finally, the detected circles in each frame are evaluated by comparing the average color in the circle with the average color of a template of the ball. This allows yielding a best match for each frame within each camera view.

In [7], the position and bounding box of ball are manually initialized at the starting time.

If a player is running near the ball, the player is marked "has ball". After that ball tracking has been stopped and the ball is searched around the player who has the ball. If the ball is found, ball tracking continues.

The flowchart of the system is composed of two alternate procedures in [8]: Ball detection and ball tracking.

In ball detection procedure, color, shape and size are used to extract ball candidates in each frame. The basic idea of ball detection is to use graph to hold multiple hypotheses of the ball's locations. There are three basic steps for this procedure: Ball candidates Detection, graph construction and ball's path extraction. The Viterbi algorithm is applied to extract the optimal path.

Once the ball is detected, the ball tracking procedure based on Kalman filtering and template matching is started. The Kalman filter predicts the ball's location in the next frame and filters the tracking result in the current frame. Template matching is used to

obtain observation. The Kalman filter and the template are initialized using detection results.

A coarse-to-fine ball detection and condensation based tracking method is proposed in [9]. Taking the similarity of weighted histogram of object region as observation model, the condensation algorithm is utilized in ball tracking. The framework of the proposed algorithm consists of three major processing modules: 1) soccer ball detection; 2) object tracking; 3) object region optimization, updating and confidence measure. The soccer ball detection module is made up of three major parts: Field extraction, shape analysis (A coarse-to-fine search strategy is used to identify a unique ball and contour of each object is traced and used to shape analyzed) and color analysis of regions and probabilities fusion (The final region similarity is fused by color and shape similarity, which should be normalized prior to fusion). The condensation algorithm (based on factored sampling but extended to apply iteratively to successive images in a sequence) is utilized to ball tracking in this method. It treats the motion image sequence as a dynamic system.

2.2.6 Trajectory association and fusion

The reconstruction of global trajectories across the multiple points of view requires the fusion of multiple simultaneous views of the same object as well as the fusion of trajectory fragments captured by individual sensors. The trajectory data from individual sensors may contain errors and inaccuracies caused by noise, objects re-entrances and occlusions.

There are two main approaches for generating the fused trajectories: fuse and then track [11], and track and then fuse [5][6][12][13][14].

A multi-camera multi-target track-before-detect (TBD) particle filter that uses mean-shift clustering is presented in [11]. A sample projection of detection mask is shown in figure 2-6.

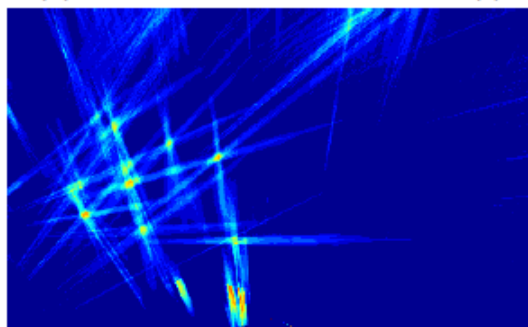


Figure 2-6: sample projection of detection mask from multiple cameras to a top view

The information from multiple cameras is first fused to obtain a detection volume using a multi-layer homography. To track multiple objects in the detection volume, unlike traditional tracking algorithms that incorporate the measurement at the likelihood level, TBD uses the signal intensity in the state representation. Moreover, as different targets can have different signal intensities on the detection volume, it is accounted for this variation in the weight update strategy. Finally, unlike traditional methods that use K-means or Mixture of Gaussian (MoG) clustering on the detections, the proposed approach does not require manual initialization of the targets or the prior knowledge of the number of clusters as it uses mean-shift on the particles, after the update step.

The association matrix also can be decided according to the Mahalanobis distance[5] between the measurement coordinates and the track prediction.

To solve the correspondence problem, a graph based algorithm can be applied [6]. The best set of tracks is computed by finding the maximum weight path. This step can be performed using the algorithm by Hopcroft and Karp.

In [12], after obtaining the transformed trajectories, the next step is to compute their relative pair-wise similarities for association and fusion in order to have a single trajectory corresponding to an object across the entire field. The following assumption is made for association and fusion: i) two trajectories that are close in space and time and ii) having similar shape are considered to be generated from the same object observed by two cameras. Two cameras mounted at different positions and having different orientations may force an affine transformation onto the object trajectory.

The parameters used for the association in [13] are: shape, length, average target velocity, sharpness of turns (which defines the statistical directional characteristics of a trajectory), trajectory mean and the PCA component analysis. With all those parameters the parameter vector is generated and the cross correlation is used as proximity measure. For the trajectory fusion, an adaptive weighting method is used, where the weights are calculated as function of the number of observations for each trajectory.

Multi camera analysis algorithms that generate global trajectories may be separated in three main classes, namely: appearance based, geometry based and hybrid approaches. Following [12], the main ideas of each class with some examples for each one are described below.

2.2.6.1 Appearance-based approaches

Appearance-based approaches use color to match objects across cameras. Color information and joint probability data association filter are used in [15] for a tracking approach of football players. A colored-based object tracking approach with a particle filter implementation is used in [16] in a multi-camera environment. TV-broadcasted images and a tracking method based on template matching and on histogram back-projection are used in [7] to solve the occlusion problem. Appearance-based methods generally suffer from illumination variations that undermine color effectiveness as a cue. Also, color information alone does not suffice to disambiguate elements of a group such as members of a team in a sport scene.

2.2.6.2 Geometry-based approaches

Geometry-based approaches establish correspondences between objects appearing simultaneously in different views. These approaches generally exploit epipolar geometry, homography and camera calibration. A method that rectifies trajectories and models people's paths is proposed in [17]. Using a non-linear approach, cameras are calibrated during an unsupervised training and trajectories are rectified. Prototype path models are built from these trajectories and a similarity measure is used to match input trajectories to path models. Eight cameras covering the penalty area are used and the tracking data of football players from multiple cameras is integrated in [18], by using homography and a virtual ground image. An extension of point distribution models is proposed in [19] to analyze object motion in their temporal, spatial and spatio-temporal dimensions. Motions are expressed in terms of modes and associated to particular behavior. Methods based on pure geometric constraints heavily rely on the accuracy during the correspondence process. For example, epipolar geometry suffers from ambiguity generated by the point-to-line correspondence [20].

2.2.6.3 Hybrid approaches

Hybrid methods use multiple features to integrate the information in the camera network. A statistical approach to associate trajectories across multiple views from airborne cameras is presented in [14]. The availability of a ground plane is assumed and a minimum duration during which at least one object is observed by two cameras. Taking as input time-stamped

trajectories from each view, the algorithm estimates the inter-camera transformations, the objects association across views and the canonical trajectories. These are considered to be the best estimates in the maximum likelihood sense. In [21], a multi-camera tracking algorithm that uses a graph representation to find the position of the football players on the pitch is presented. The integration of multiple features from multiple views is proposed in [22] (e.g texture, color, region, motion). The trajectories are updated by back-projecting the 2D observations of the features and weighting them adaptively to their self-evaluated reliability. An algorithm that combines particle filtering and belief propagation in a unified framework is presented in [23]. Local particle filtering trackers interact with each other via belief propagation and compensate for poor individual observations. This algorithm is restricted to overlapping areas and relatively short time duration of occlusions.

2.2.7 Overview of techniques

The following table presents an overview of all the reviewed sport video analysis techniques:

Reference	Sport	Background extraction	Foreground analysis and blob detection	Team identification	Player tracking	Ball tracking	Trajectory association and fusion
[2]	Football	Statistical adaptive methods, Gaussian distribution	Graph representation	Vertical intensity distribution	Optimal path in the graph	-	-
[3]	Football	Learning histogram	Connected component analysis	-	-	Roundness, eccentricity and color	-
[4]	Football	-	Morphological filters, connected component analysis	Color histogram contrast	-	Hough transform	-
[5]	Football	Mixture of Gaussians	Measurement covariance	Histogram intersection	Kalman filter	-	Mahalanobis distance
[6]	Football	Background with updating procedure	Connectivity analysis	-	Maximum a Posteriori Probability	-	Graph based algorithm, Hopcroft and Karp algorithm
[7]	Football	Peak values of histogram, morphological filtering	-	Vertical distribution of colors	Template matching, Kalman filter	Template matching, Kalman filter, search windows	-
[8]	Football	-	-	-	-	Color, shape, size, Viterbi algorithm, Kalman filter, template matching	-
[9]	Football	Mean color, color distance	-	-	-	Weighted histogram, condensation algorithm	-
[10]	Football	-	-	-	Template matching	-	-
[11]	Basketball	-	-	-	-	-	Particle filter, multi-layer homography, K-means, mixture of Gaussian clustering
[12]	Football	-	-	-	-	-	Feature vector correlation
[13]	Basketball and football	-	-	-	-	-	Feature vector correlation
[14]	-	-	-	-	-	-	Graph based algorithm

Table 2-1: Overview of sport video analysis techniques

2.3 Content sets

In order to develop and test the algorithms and methods to build the sports video detection and tracking system, to have content sets containing ideally not only multicamera video sequences but also ground-truth data or calibration data is key.

All the multicamera sports video content sets used in the project have been recorded with static cameras. Therefore, the background view is the same for each sequence of each camera, allowing to apply a constant homography.

2.3.1 ISSIA Soccer Dataset

The public ISSIA dataset¹ is composed of:

- Six synchronized views acquired by six Full-HD cameras, three for each major side of the playing-field, at 25 fps (6 AVI files).
- Manually annotated objects position of two minutes of the game. These metadata provide the positions of the players, referees, and ball in each frame of each camera (6 XML files). The players have the same labels while they move in the six views that correspond to the numbers on their uniforms. The player labels of the first team start from 1 while the player labels of the second teams start from 201.
The first 300 frames of each sequence have not been labeled in order to provide an initial phase to initialize the background subtraction algorithms.
- Calibration data: Pictures containing some reference point in to the playing- field and the relative measures for calibrating each camera into a common world coordinate system (6 pdf files).

All cameras are DALSA 25-2M30 cameras.

The positions of the six cameras on the two sides of the field are shown in figure 2-7.

In figure 2-8, an example of each video is shown. Those examples allow to know where are placed the cameras and the point of view of each camera.

¹ <http://www.issia.cnr.it/htdocs%20nuovo/progetti/bari/soccerdataset.html>

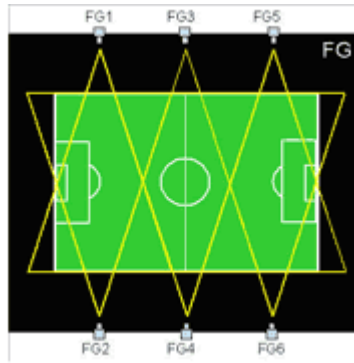


Figure 2-7: Positions of the six cameras of ISSIA Soccer Dataset



Figure 2-8: Examples of the different videos of ISSIA Soccer Dataset

2.3.2 TRICTRAC Test Case Scenarios

TRICTRAC is an innovative 3-year project in the field of image processing. The goal is the development of algorithms for the tracking of objects in real time in one or more live video streams. TRICTRAC is a joint project between the INTELSIG group at the Montefiore Institute of the Université de Liège, the tele group at the Université Catholique de Louvain and the research center Multitel.

For the TRICTRAC project some video clips were rendered. For the first rendering campaign, 3 scenarios were tested with several cameras.

For the dataset², 5 different virtual cameras have been used. The camera 9 is used with two different fields of view. Some examples of these videos are shown in figure 2-9.



Figure 2-9: Examples of different videos of TRICTRAC Test Case Scenarios

2.3.3 PETS Football Dataset

The PETS football dataset³ consists of football players moving around a pitch. There are some videos (training data) intended for developing background models of the scene. The sequences have been specifically chosen to contain the least amount of player action.

The annotation (ground truth) for camera view 3 is available, with details of the format. An AVI file of the ground truth for camera view 3 is also available. The annotation for camera views 4 and 5 are not available yet.

² <http://www.multitel.be/trictrac/?mod=3>

³ <http://www.cvg.cs.rdg.ac.uk/VSPETS/vspets-db.html>

The coordinate space (plan of camera locations) is shown in figure 2-10. Details on camera calibration are available also in the dataset files. In figure 2-11, an example of each video is shown.

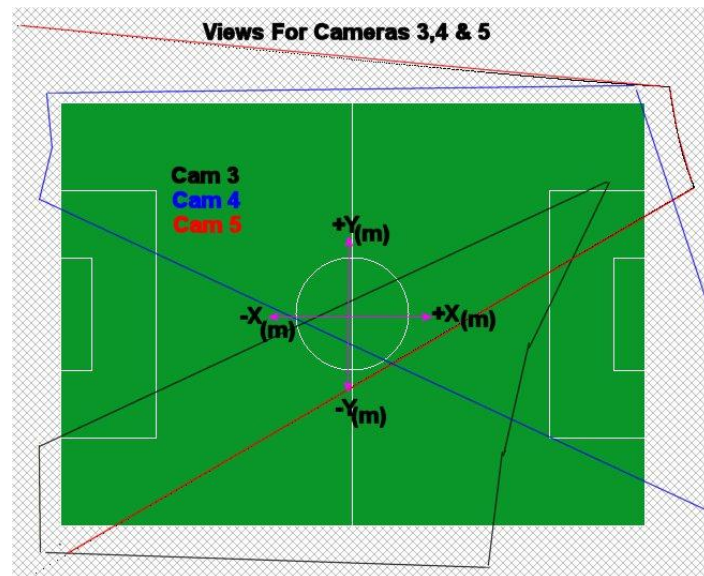


Figure 2-10: Positions of the cameras of PETS Football Dataset

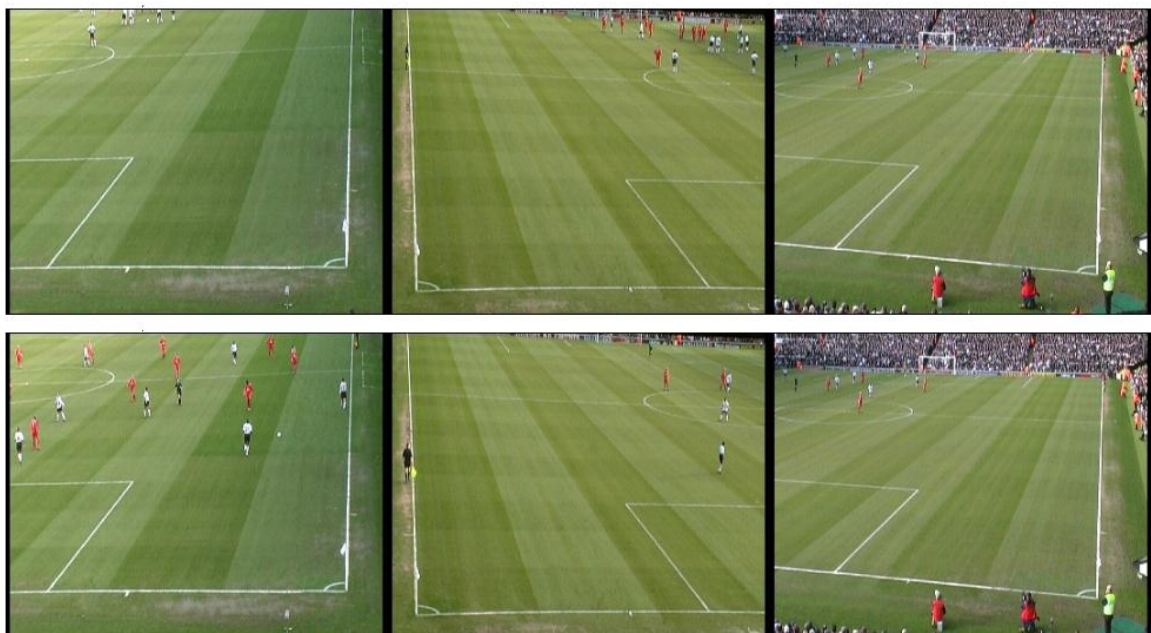


Figure 2-11: Examples of the different videos of PETS Football Dataset. Training up and testing down

2.3.4 APIDIS Basketball Dataset

The basketball dataset⁴ has been acquired beginning of April 2008 in Namur, Belgium. The game which has been recorded was the final of the female Belgian basketball league. The dataset is composed of a basketball game:

- Seven 2-Mpixels color cameras around and on top of a basketball court
- Time stamp for each frame. All cameras being captured by a unique server at ~22 fps.
- Manually annotated basketball events for the entire game. These metadata provide events relative to the basketball game, e.g. ball possession periods, throws and violations.
- Manually annotated objects positions for one minute of the game. These metadata provide the positions of the players, referees, baskets and ball in each frame of each camera.
- Calibration data. These are measures of the basketball court and calibration pictures that can be used for calibrating each camera into a common world coordinate system.

Sample pictures are shown in figure 2-12 for each camera. The configuration of the video sensors is shown in figure 2-13.

⁴ <http://www.apidis.org/Dataset>

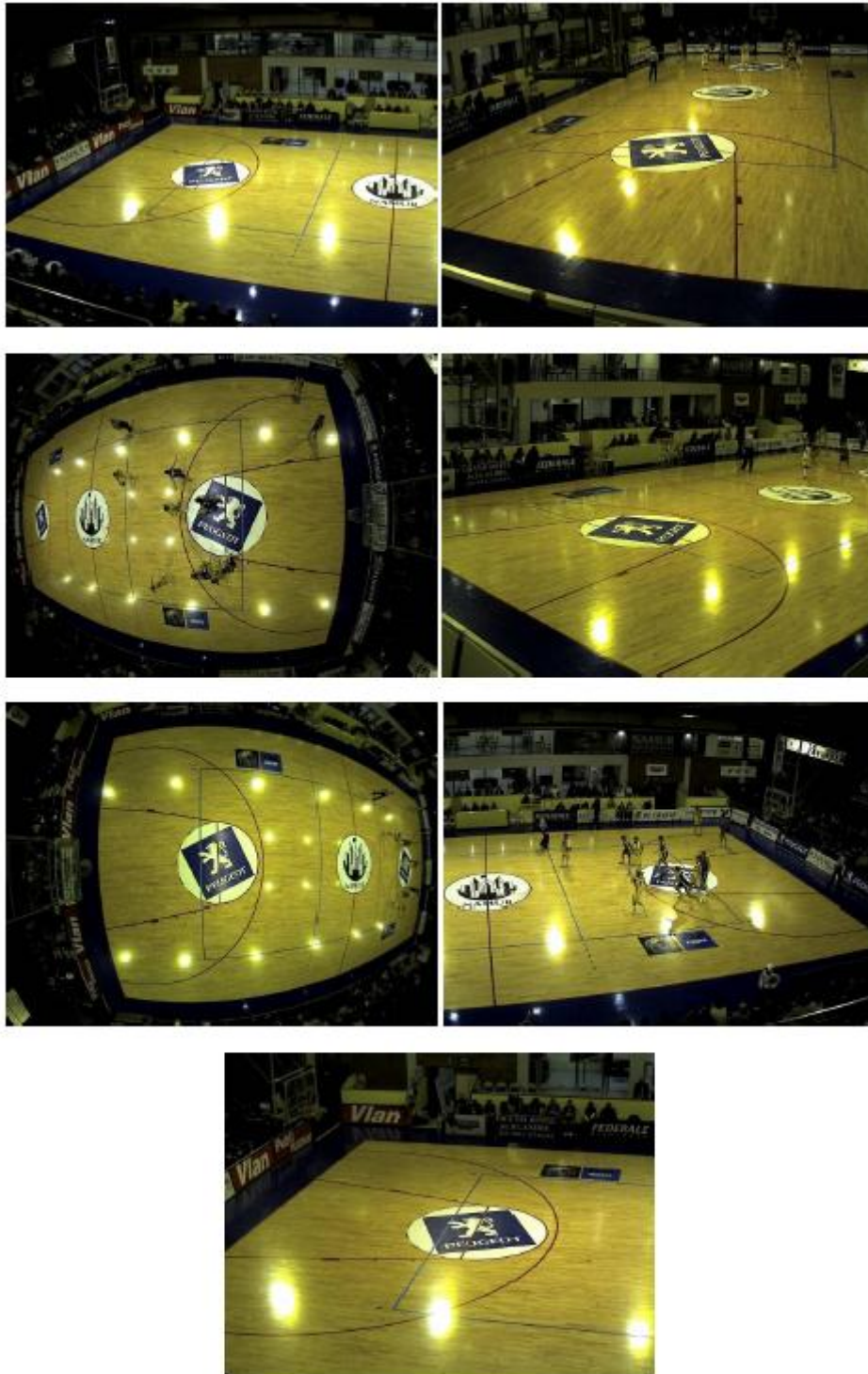


Figure 2-12: Examples of the different videos of APIDIS basketball dataset

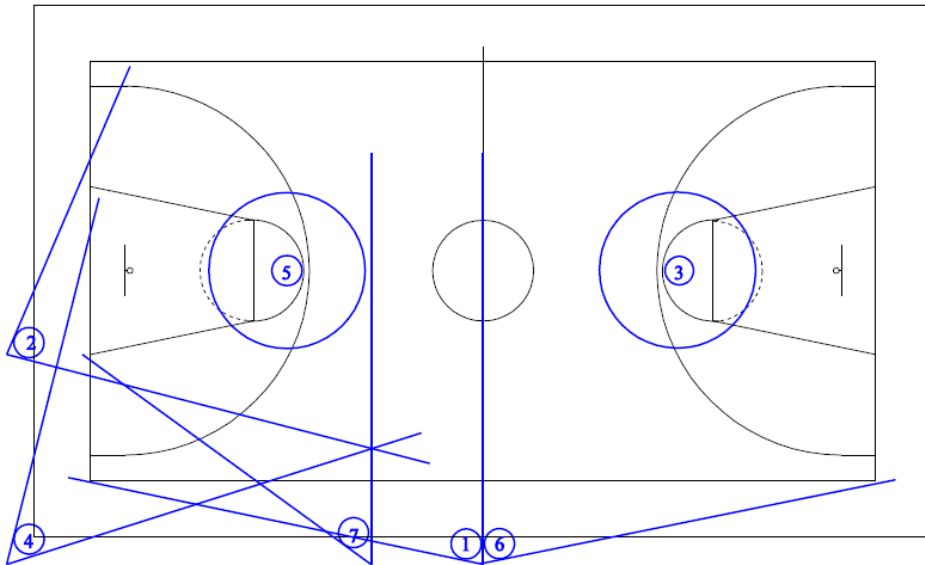




Figure 2-13: Positions of the cameras of APIDIS basketball dataset

2.3.5 3DLife ACM Multimedia Grand Challenge 2010 Dataset

The dataset⁵ features video from 9 CCTV-like cameras placed at different points around the entire court (see figure 2-14). In addition, audio from 7 on the 9 cameras (each camera with the  microphone symbol in the top left of the image) is also available. Videos are ASF files and encoded using an MPEG-4 codec. 7 of the videos (taken from the cameras with the  microphone symbol in the top left of the image) are recorded with a resolution of 640x480 pixels from Axis 212PTZ network cameras. The two other cameras have a resolution of 704x576 pixels and are captured using Axis 215PTZ network cameras. Although, the start time of each video is synchronized via software at the start of each sequence, the videos are not genlocked together.

⁵ http://www.cdvp.dcu.ie/tennisireland/TennisVideos/acm_mm_3dlife_grand_challenge/

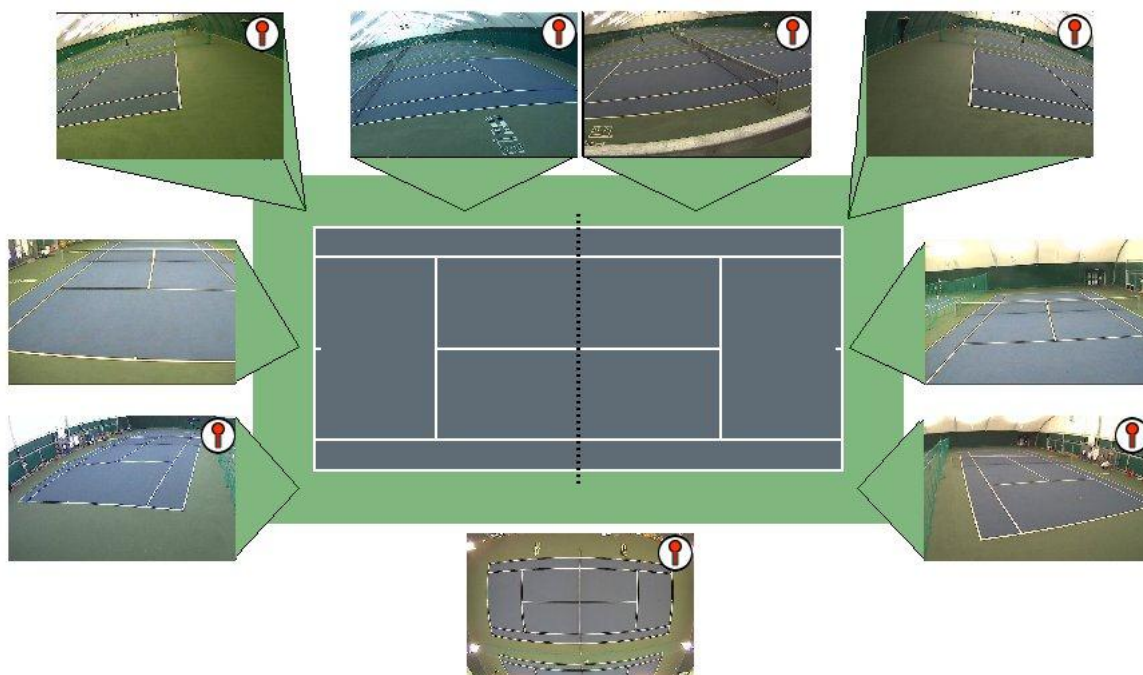


Figure 2-14: Positions and examples of the different videos of 3DLife dataset

2.3.6 CVBASE '06 Dataset

To encourage research on computer vision in sport environments, the organizers of CVBASE 2006 provide a dataset⁶, which is available for download from their page.

The CVBASE 06 Dataset consists of three types of data:

- Video data (.avi, DivX compressed)
- Annotations (individual player actions, group activity). Suitable for use as a gold standard.
- Trajectories (player positions in court and camera coordinate systems). These are not intended to be used as a gold standard, since their accuracy is not particularly high.

Dataset includes three types of sports:

- Basketball (videos only). Figure 2-15.
- Squash (videos, trajectories, annotations). Figure 2-16.
- European (team) handball (videos, trajectories, annotations). Figure 2-17.

⁶ <http://vision.fe.uni-lj.si/cvbase06/downloads.html>

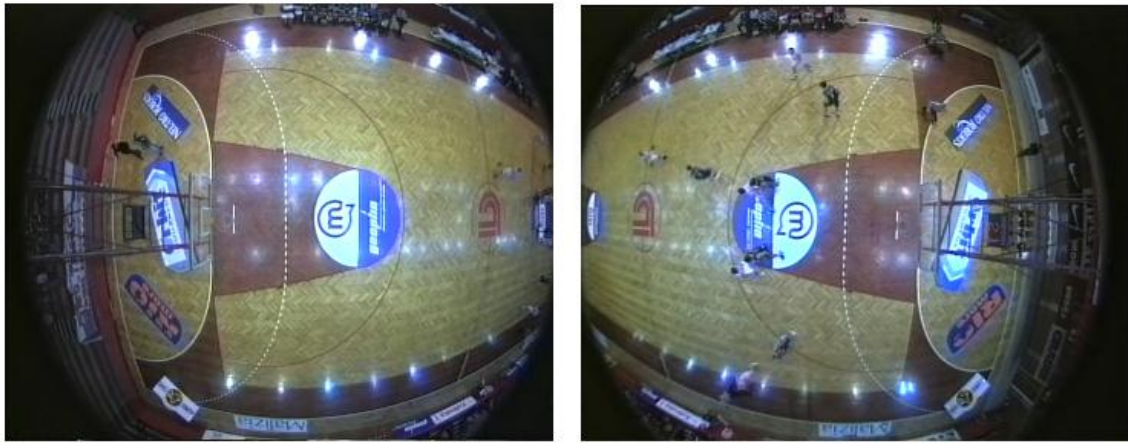


Figure 2-15: Examples of the different videos of CVBASE basketball dataset



Figure 2-16: Examples of the different videos of CVBASE squash dataset

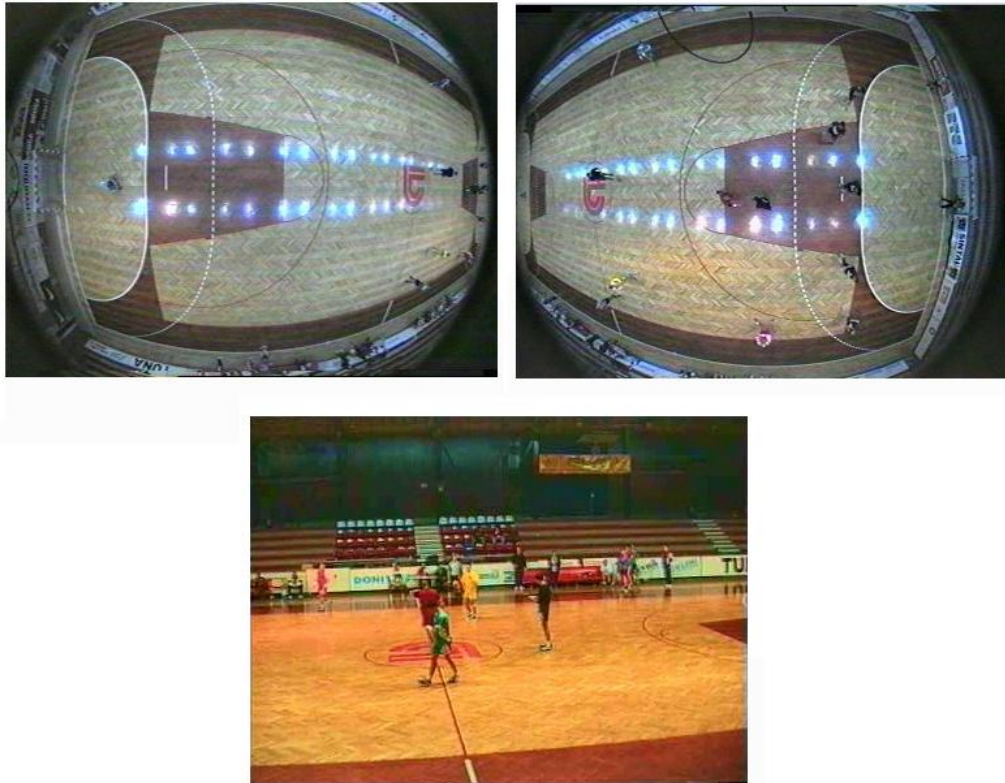


Figure 2-17: Examples of the different videos of CVBASE handball dataset

2.3.7 Overview of content sets

In the following table an overview of all the content sets is presented:

Content Set	Sport	Number of cameras	Calibration	Ground Truth	Others
ISSIA	Soccer	6	Pictures with reference points	Yes	Camera datasheet
TRICTRAC	Soccer	5 (virtual)	No	Yes	-
PETS	Soccer	5 (3 available)	Yes but not available	Only for camera 3	Training and testing sequences
APIDIS	Basketball	7	Measures of the basket ball court and calibration pictures	Yes	Camera and lenses datasheet, annotated basketball events
3DLife	Tennis	9	Images for camera intrinsic parameters and for court calibration	No	Audio from 7 cameras, data from inertial measurement units
CVBASE	Basketball Squash Handball	2(basketball) 1(squash) 3(handball)	No	Yes (squash and handball)	Matlab interface for easy access to video data

Table 2-3: Overview of content sets

2.4 Commercial products⁷

In this section some commercial products are presented.

2.4.1 TRACAB Image Tracking System⁸

The TRACAB Image Tracking System™ uses this technology to identify the position of all moving objects in a sports arena in real-time. The outcome of the analysis is a data feed with X, Y and Z coordinates for all objects. The precision of the measurements is below one decimeter. The data feed is used by different applications serving various customer needs.

One of the biggest advantages is that the system uses no transmitters or any other kind of device on the players or the ball. The system does not in any way interfere with the game.

The Tracab system has been modified and developed since the first production in the Swedish Royal League Final in the spring of 2005. During over a 3,000 live productions the system has been exposed to numerous challenges in terms of weather, light conditions, stadia conditions, logistic challenges, etc.

The system consists of two multi-camera units which contains eight small cameras each. By using more than one camera the system gets less sensitive to a camera failure, even if one camera fails the unit will still be operational. It also provides the camera units with the feature of dynamic resolution, meaning that the individual cameras can be angled so that the camera unit delivers a higher resolution in the parts of the pitch where there are most crowded and occluded situations, such as the penalty boxes.

The Tracab system is based on SAAB patented stereo vision technology. This basically means that every inch of the pitch is constantly covered by at least two cameras. The x, y, z measurements for the objects on the pitch are made through analyzing both these images and, most importantly, the difference between them. This method results in true three dimensional tracking.

⁷ These descriptions have been extracted from the publicity of the products. As they are not free products, we have not been able to test them.

⁸ www.tracab.com

2.4.2 AmiscoPro⁹

The Amisco System is a solution for the creation and managing of sports analysis data. The key functionality is to measure and store tactical and fitness data for sport. The Amisco system is based on a patented “tracking” technology and relies on a set of sensors installed on the stadium and produce an animation in 2D with tactical and physical exclusive data.

This system supplies an objective measure of the positions and the movements of the players on the pitch. The used technology consists of one “passive tracking” (The players are equipped with no equipment). The data obtained is intended for AMISCOPRO interactive software for consultation.

8 sensors are implanted around the field. They capture and measure objectively the position and the movements of the players, the referees and the ball 25 times per second during the entire game.

Some of the most famous teams that use this system are Real Madrid, Valencia CF, Chelsea FC, Liverpool FC, Bayer Leverkusen and Olympique de Marseille.

2.4.3 Vis Track¹⁰

The VIS.TRACK system enables you to evaluate the performance data for all the players in real time and archive this. With two cameras, to analyze everything that happens in a match is possible, because not only the player data, but also the ball data, is captured. With the VIS.TRACK software, this information can be represented in parallel in 3-D animations and graphs. A permanent installation is not needed for the VIS.TRACK system. The system can be set up in just 30 minutes and can therefore be used not just in the stadium but also at the training ground.

Some of the functions that the application provides are: all players and ball tracking in real time, players trace, real time speed and distance, compactness of your team, 3D-Heat map of your team, real time speed measurement, performance overview.

The tactic window provides information about running performance and the tactical behavior of the team. For example: players can be linked among each other or with opponents showing the distance between them. You are able to choose the point of view in all the animations, which provides you with a valuable insight of the events on the pitch.

⁹ <http://www.sport-universal.com>

¹⁰ <http://www.cairos.com/unternehmen/vistrack.php>

This window can be equipped with several independently chosen diagrams. Besides the live speed of a player, the covered distance and the speed sectors, the diagrams are also showing team values and player comparisons. Individual threshold values can be defined and analyzed in real time.

Goals, shot on goal, crosses, corner kicks, duels and further actions are linked to the video time code and can be displayed with a mouse click. To display single scenes and rejoin them to a highlight video afterwards is possible. The statistic window provides a good overview about all statistical data of the match and offers the possibility to compare players.

3 Base system and common modules

3.1 Introduction

This section describes the base system, the modifications applied on it and the new modules created. All the contents of the sections are used in the systems of individual sports and team sports.

3.2 Base system overview¹¹

The base system used in this work starts from a video analysis system for abandoned and stolen object detection [24]. This system is designed to work as part of a video-surveillance framework capable of triggering alarms for detected events in real time. This requirement imposes limits on the time complexity of the algorithms used in each of the analysis modules. The system's block diagram is depicted in Figure 3-1.

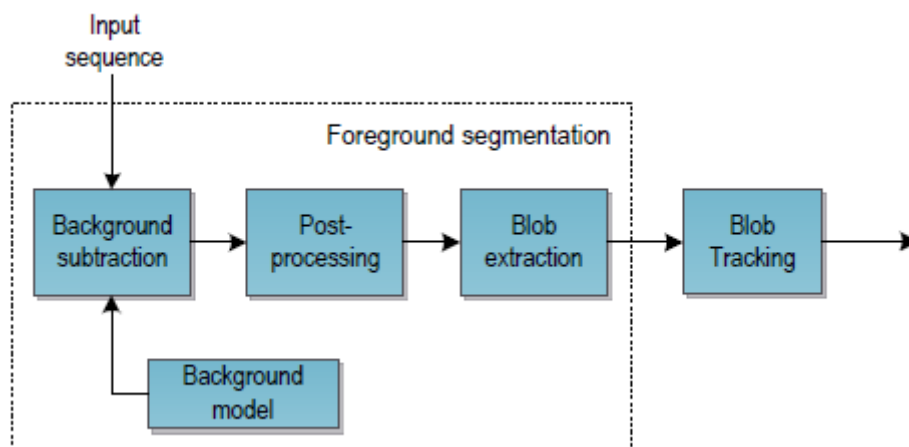


Figure 3-1: Block diagram of the base system[24]

After the initial frame acquisition stage, a foreground mask is generated for each incoming frame at the Foreground Segmentation Module. This foreground mask consists on a binary image that identifies the pixels that belong to moving or stationary blobs. Then, post-processing techniques are applied to this foreground mask in order to remove noisy artifacts and shadows. After that, the Blob Extraction Module determines the connected

¹¹ This overview is extracted from [24].

components of the foreground mask. In the following stage, Blob Tracking Module tries to associate a unique ID for each extracted blob across the frame sequence.

The base system is described with greater detail in annex A.

3.3 Base system modifications

In this section, the modifications applied to the base system are described. There are five main modifications:

- Generation of background
- Extraction of the base mid-point
- Generation of the file with coordinates (for further fusion)
- Adjustment of parameters to the characteristics of the videos
- Addition of a mask for limiting the tracking area

Applying the modifications, an acceptable tracking is obtained for the individual cameras. After it, the fusion process (described in the next chapters) uses the obtained tracking information for applying the fusion and generating the global trajectories across the field.

3.3.1 Generation of background

The tracking system used needs an initial background image of the scene for each video analyzed. To get this image two simple methods, which provide similar results, have been developed:

- Junction of portions from different frames
- Mode of pixels from a set of frames

These methods are necessary when an image of the background without players is not available. In video-surveillance more complex methods are used for hot start up (to start without initialization) or with updating background, but this method covers quickly and easily the objective.

In live applications, players usually warm up some minutes before the game. These sequences can be used for system initialization, including among others, background generation.

In the systems designed both methods have been used: for the individual sports system both methods have been used, whilst for the team sports system only the second method has been used.

3.3.1.1 Junction of portion from different frames

This supervised method is based on finding some portion in frames without players in the field. Once enough frames to compose a full background are found, the final image is created.

The portions used in this project are halves, but smaller portions can be taken, such as quarter-frame or vertical strips less than half frame

The (non assisted) procedure used is: First, the video frames are watched using a video player trying to locate regions where there are no players. Then a first approximation of the background is generated with the frames located in the previous step. If a region contains background generated players (or part), the frame number chosen for that region can be increased or decreased to get an empty region, and the background is generated again. If players appear completely, another frame should be chosen by visualizing the video again.

The advantage of this method is the reduced computational cost.

The disadvantages of this method are that sometimes finding frames without players is not easy and that it is a supervised method.

An example of incorrect background with part of a player is shown in figure 3-2. An example of correct background is shown in figure 3-3.



Figure 3-2: Example of incorrect background



Figure 3-3: Example of correct background

3.3.1.2 Mode of pixels from a set of frames

This method is based on choosing (watching the video using a video player as in the case above) a set of frames and generating each pixel of the background with the mode of all the pixels in the position of the pixel generated.

Other operations like mean or median have been tested with worse results.

The advantage of this method is that it needs less supervision than the other method. The script requests an initial frame and a number of frames to analyze. The original frame is chosen after watching video display frames with few players and without static players, which may cause problems when generating the background. By increasing the number of frames processed improves the results, increasing the computational cost.

The disadvantage of this method is the high computational cost due to mode operation.

Using this method sometimes small areas with strange colors appear caused by any person or object moving slowly in the set of chosen frames to generate the background. One example of this area is shown in figure 3-4. This colored area is caused by a person moving behind the advertising. These areas generally do not cause malfunction because the erosion and reconstruction operations make them disappear from tracking.



Figure 3-4: Example (detail) of malfunction in background generator

An example of incorrect background is shown in figure 3-5. An example of correct background is shown in figure 3-6.



Figure 3-5: Example of incorrect background



Figure 3-6: Example of correct background

3.3.2 Extraction of the base mid-point

The base system obtains two points for each player in each frame, defining the bounding box of the player (p1 and p2). Each point has its coordinates x and y: p1 corresponds to the upper left point of the bounding box and p2 corresponds with the lower right point of the bounding box. In figure 3-7 some examples of bounding box and base mid points are shown to clarify.

Base mid-point (b) coordinates (x_b and y_b) are defined as follows:

$$x_b = x_{p1} + \frac{x_{p2} - x_{p1}}{2}$$

$$y_b = y_{p2}$$

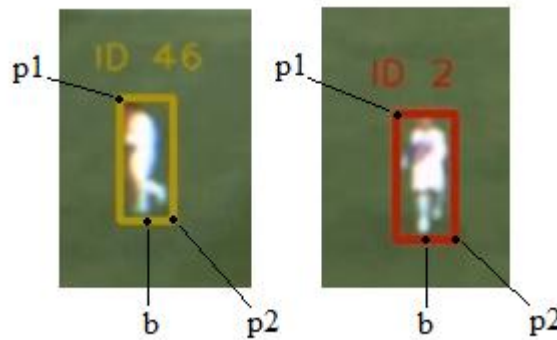


Figure 3-7: Examples of bounding box (yellow and red) indicating the base mid point (b) and points p1 and p2

The base mid-point is taken because when the homography is applied, the correspondence is calculated for points on the floor of the field. The base point is closer to the floor and presents a minor error than selecting for example the center point on the bounding box (which is typically used in other tracking applications).

3.3.3 Generation of the file with coordinates

This modification is necessary because the original system does not generate a file with the coordinates of each blob. A video with the bounding boxes of the blobs tracked added to the original video is generated when the original base system is executed.

With this modification, when a video is processed by the base system, an annotation file in text format is generated with the following format for each line:

frame blobID x y

Where *frame* is the number of the frame annotated, *blobID* is the blob identifier annotated, and *x* and *y* are the coordinates of the base point of the bounding box of the player. An example of the format is shown in figure 3-8:

```
8 1 295 176
8 2 382 208
8 3 780 218
8 4 91 248
8 5 431 274
8 6 255 289
8 7 358 315
8 8 57 323
8 9 122 331
8 10 292 372
```

Figure 3-8: Example of some lines of the generated file in tracking

In figure 3-8, each line of the file describes the coordinates *x* and *y* from the blobs 1 to 10 in frame 8.

3.3.4 Adjustment of parameters to the characteristics of the videos

The original configuration of the base system was inadequate for sports. The main problem found was that the system confused the different players on a team. When a player tracking was lost, the system took the tracking of another player even if the new detected player was far away from the original tracked player.

The base system is a mono-camera system for video surveillance, where usually there are not so many blobs tracked simultaneously as in the case of multiplayer video. As there is only one camera, the system tries to track the person during all the video. With multi camera fusion, even if a person is tracked with different blobs, the final trajectory can track the player during all the video in a single trajectory.

The main parameters of the base system are:

- **MIN_ACTIVE**: Controls the number of initial matches to consider a blob stable over time.
- **MAX_NO_MATCHING**: Number of consecutive frames of no-matching to remove the blob.
- **THRESHOLD_H**: Threshold of height difference.

- THRESHOLD_W: Threshold of width difference.
- THRESHOLD_DIST: Threshold of color difference.
- RADIO: Basic value of maximum distance of search.
- MAX_ID: The highest value of the ID.
- IN0, IN1, IN2, IN3: Constant values of random color¹².
- MAXOBJECT: Constant values of max number of objects.
- MAXINFO: Constant values of max number of info.

The most interesting parameters to modify are: MIN_ACTIVE, MAX_NO_MATCHING and RADIO. For example, reducing RADIO parameter, the problem of continuing the tracking of a player with the tracking of another player is reduced.

After some testing and parameter variations, the final values for both systems designed (individual sports and team sports) are:

- MIN_ACTIVE_2 = 10
- MAX_NO_MATCHING_2 = 5
- THRESHOLD_H_2 = 30
- THRESHOLD_W_2 = 30
- THRESHOLD_DIST_2 = 30
- RADIO_2 = 30.0
- MAX_ID_2 = 60000
- IN0_2 = 0;
- IN1_2 = 85
- IN2_2 = 170
- IN3_2 = 255
- MAXOBJECT_2 = 10
- MAXINFO_2 = 100

¹² These values define the random color of the bounding boxes of the tracked players, shown in the output video generated by the tracking system.

3.3.5 Addition of a mask for limiting the tracking area

In some videos (especially in the ISSIA soccer dataset), there are people moving out of the field (e.g., coach, auxiliary referee) and dynamic objects (e.g. advertising) that cause errors in tracking. To solve these errors, a mask is manually generated for each video to avoid tracking in the areas out of the field.

The procedure used is: First, A mask is defined with all its pixels with value 1. After that, the video frames are watched, trying to locate regions where problems may occur, such as dynamic advertisements or TV cameras. Then, some points are selected in the background image with the Matlab data cursor and annotated. Joining these points with lines, a region to be removed from the analysis is defined. All pixels in this region take values of 0. If necessary, more than one region can be defined.

The mask is applied with a logical AND operation to the extracted foreground in each frame. Some examples of the generated masks are shown in annex D.

3.4 Homographies

A homography is an invertible transformation from a projective space (for example, the real projective plane) to itself that maps straight lines to straight lines and points to points.

Formally, a projective transformation in a plane is a transformation used in projective geometry: it is the composition of a pair of perspective projections. It describes what happens to the perceived positions of observed objects when the point of view of the observer changes. Projective transformations do not preserve sizes or angles.

As presented in [6] and [12], given a point to point correspondence between two views (image plane to top-view), the homography can be estimated using:

$$\psi' = H\psi$$

Where $\psi = (x; y; 1)$ is a point in one view and $\psi' = (x'; y'; 1)$ the corresponding point in the second view. Equation can be expressed using homogeneous coordinates:

$$x' = \frac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + h_{33}}$$

$$y' = \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}}$$

Where h_{ij} are entries of the matrix H :

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix}$$

The homographies H_k ($k=1, 2, \dots, N$) are estimated to generate the transformations that are used to correspond each camera image plane to the ground plane top view. Each projected local trajectory is modeled as polygonal line in $2D + t$.

The projected trajectories of a given object, simultaneously viewed by two different cameras can result in non-coincident trajectories on ground plane top view.

The code used is public¹³ and follows the normalized direct linear transformation algorithm given by Hartley and Zisserman[20].

In figure 3-9, an example of the result is shown. Note that in the system the homography is applied to a single point (the base mid-point –see section 3.3.2-) for each player tracked. The top view image (b) is presented for facilitate understanding the homography result. The red points in the figure indicate the points used for generating the homography matrix H .

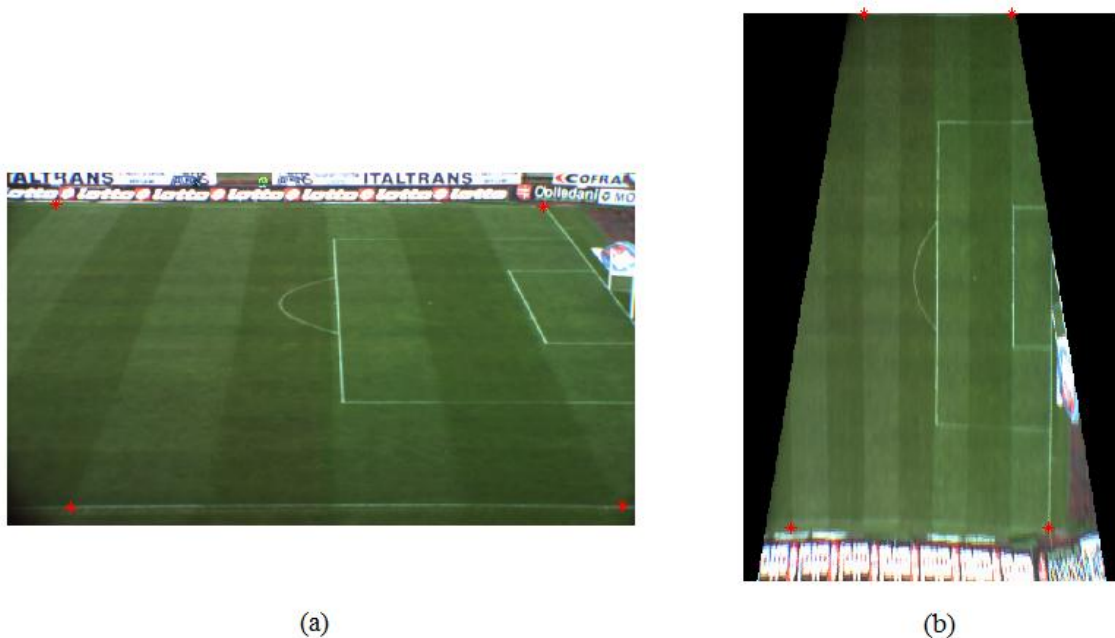


Figure 3-9: Example of homography: (a) image plane, (b) Top view

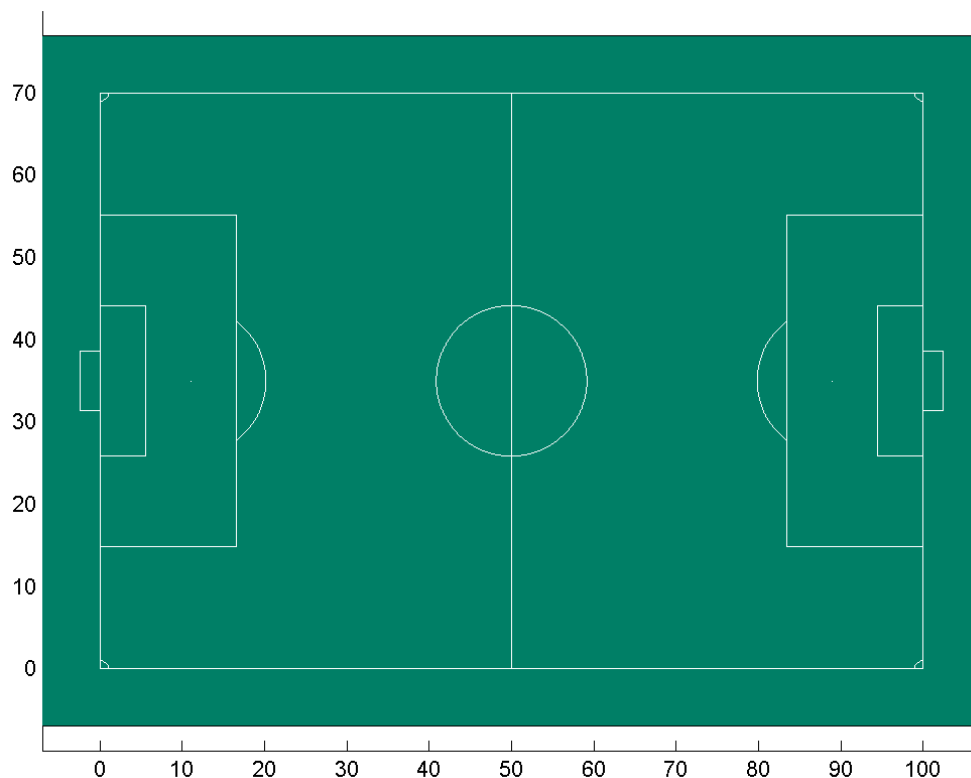
¹³ <http://www.csse.uwa.edu.au/~pk/research/matlabfns/>

Homographies belong to the system initialization, after background generation described in section 3.3.1.

3.5 Field representation

For the representation of the individual or fused trajectories, a set of field representation has been designed. The fields represented are for the following sports: tennis, basketball and football. Each one of them has been developed in a different Matlab script. Each script draws lines according to the dimensions of the field. The dimensions used are the standard dimensions of each sport¹⁴, but can be easily changed in the initialization module if the filmed field has specific dimensions (generally there are some distances in standard dimensions which can take a range of values instead of a constant dimension).

The resulting fields of each script are shown in the following figures.



¹⁴ Tennis: <http://es.wikipedia.org/wiki/Tenis>

Football: <http://es.wikipedia.org/wiki/F%C3%BAtbol>;

Basketball: <http://es.wikipedia.org/wiki/Baloncesto>

Figure 3-10: Football field

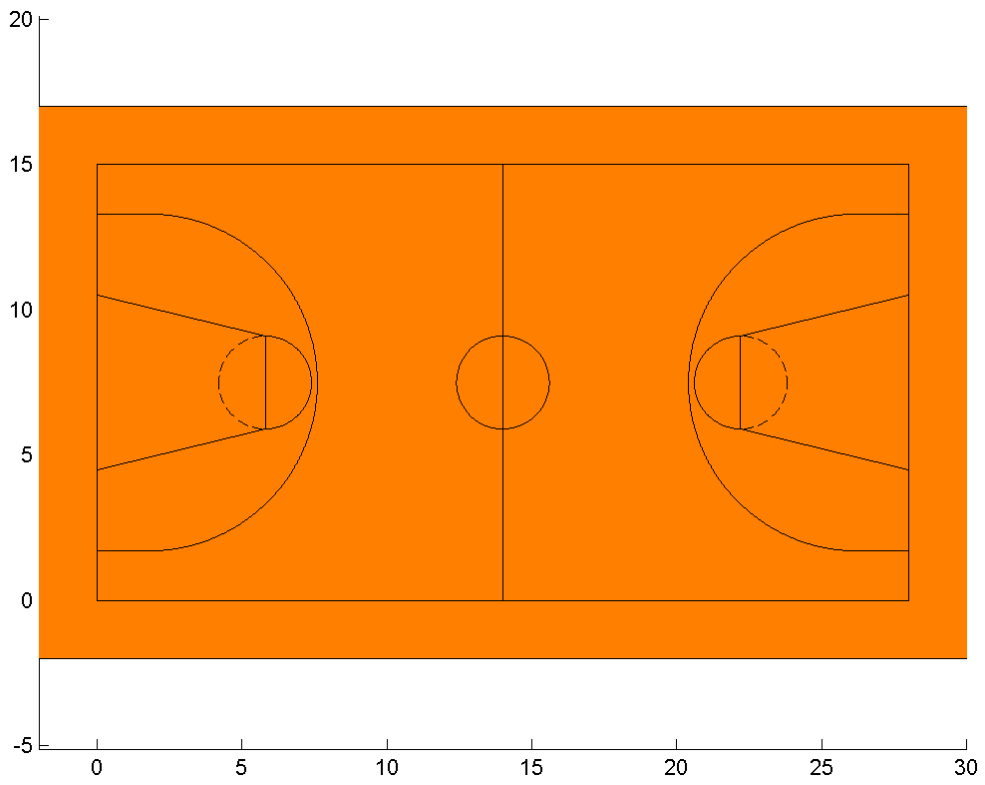


Figure 3-11: Basketball field

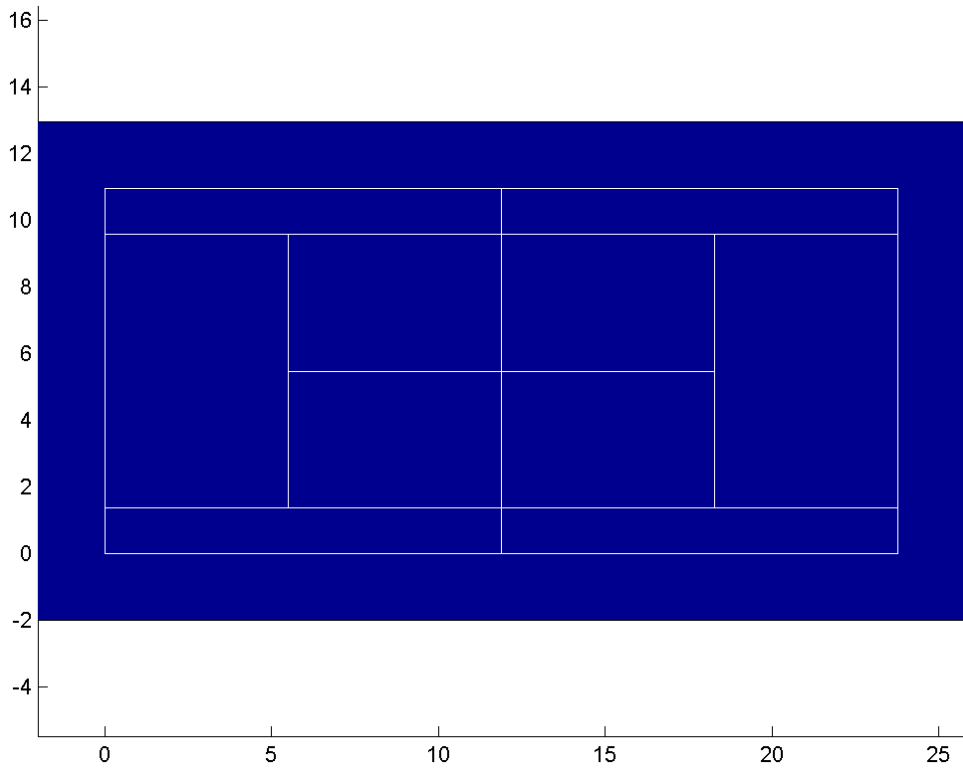


Figure 3-12: Tennis field

4 Individual sports

4.1 Introduction

The first system presented is centered on individual sports. We refer individual sports for sports where every player has his own area of the field and no other player or referee can enter in the area of the monitored player. This is a fundamental aspect because the fusion step is based on this characteristic.

Some examples of this kind of sports are tennis or paddle tennis.

4.2 System

In this section the global system is presented. The block diagram of the system is depicted in figure 4-1.

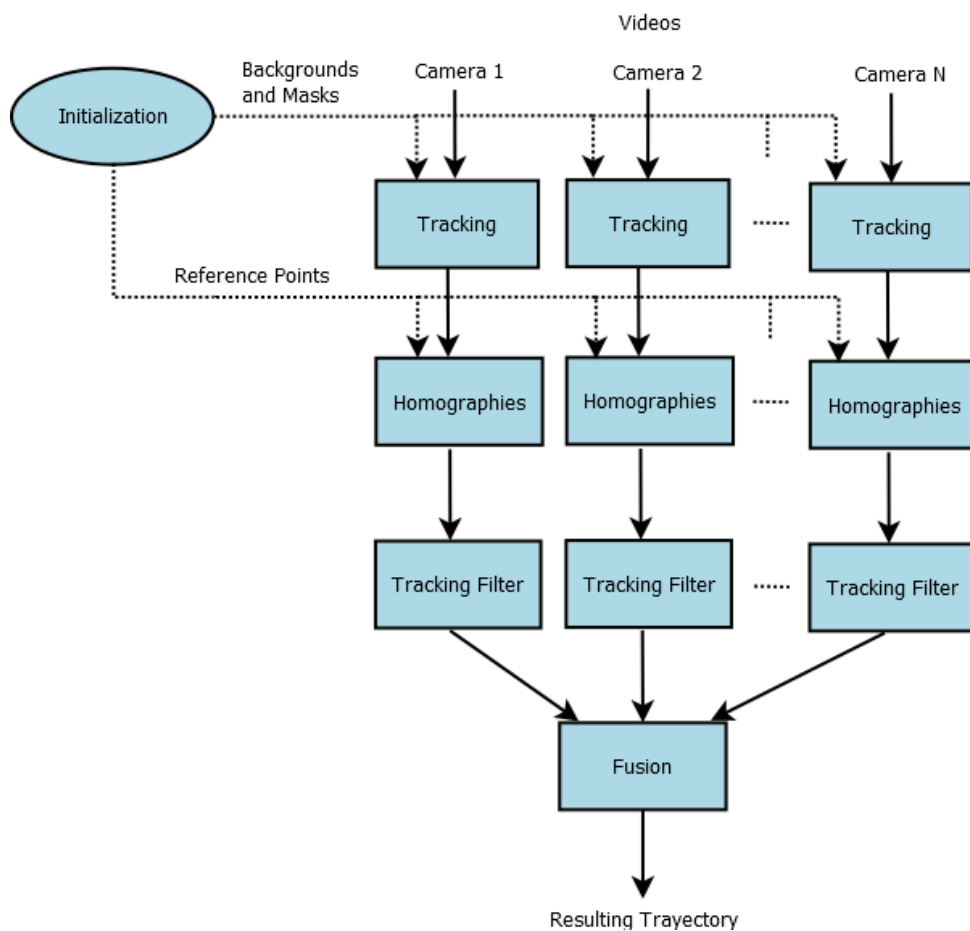


Figure 4-1: Block diagram of the system for individual sports

The block diagram is individual for each player. The videos used from the content set are those covering only the areas of each player, not global views. That is, for a full match there will be necessary two parallel systems. This condition is necessary for the characteristics of the fusion block.

The **Initialization block** is executed first. It generates the field representation, the backgrounds, the masks and the homography reference points. The generation method of each of these elements is detailed in the corresponding section. For the used videos applying masks was not necessary so empty masks were generated.

The **Tracking block** receives the videos and the masks from the different cameras which are recording the player and generates the tracking from its own perspective.

The Homographies block receives the tracking of the player from each camera and the reference points, and it generates the top view tracking from each camera.

The **Tracking filter block** receives and filters the tracking for the top view tracking from each camera. Some examples of filtering criteria are: possible regions in the field, minimum number of frames for the blob, distant blobs non belonging to the player. The block generates the top view filtered tracking from each camera.

Finally, the **Fusion block** joins the processed tracking from each camera and generates the overall trajectory along the entire field. The fusion method will be described in the next section.

4.3 Fusion

For individual sports, the fusion process is relatively simple. The main advantage is that every blob in frame t belongs to the tracked player.

$$x_t = \frac{1}{N} \sum_{i=1}^N x_{i,t}$$

$$y_t = \frac{1}{N} \sum_{i=1}^N y_{i,t}$$

Where (x_t, y_t) are the fused coordinates in frame t, $(x_{i,t}, y_{i,t})$ are the coordinates in frame t from the i-camera, and $i=1 \dots N$ are the N cameras which have a recorded blob corresponding to the player in frame t.

Additionally, a weight can be added to each blob coordinates depending on the position, height and precision of each camera.

4.4 Adjustments, testing and results¹⁵

For the integration, the used videos are from 3DLife ACM Multimedia Grand Challenge 2010 Dataset (see section 2.3.5). Specifically, the videos are from the cameras 1, 3 and 4 recording the blue player in game 1. The other videos have problems during the tracking for many causes: the net that separates the fields is moved by the wind, camera that automatically focuses and blurs, camera slightly moved by the wind, etc. Solving these problems was not in the objectives of the project and will be described as future work.

The tracking module used is the one from the base system described in annex A with the modifications described in section 3.3. The average time of the tracking processing in the tennis videos is 21.9 FPS.

The homography reference points are calculated and introduced manually. The procedure used, implemented via a Matlab script, is: After obtaining the background, it is represented in a Matlab figure. Then, the reference points are selected with the data cursor and annotated to be applied later in the homographies.

In the Tracking filter module, the criteria (heuristically set) for filtering are: Minimum of 10 frames per blob and a region in the field from approximately 70 cm from the net. This minimum distance is applied because during the video the ball moves the net and a “ghost” static blob is generated in the region where the net differs from the net modeled in the background.

The fused trajectory starts at frame 25 to allow initializing the tracking system.

In figure 4-2, the resulting trajectory from the tracking from each camera is presented, and in figure 4-3, the resulting trajectory from fusion is shown.

¹⁵ A web page has been created to add videos and results, where a video with the result of the system has been added.

<http://www-vpu.eps.uam.es/publications/DetectionAndTrackingInMulticameraSportsVideo/>

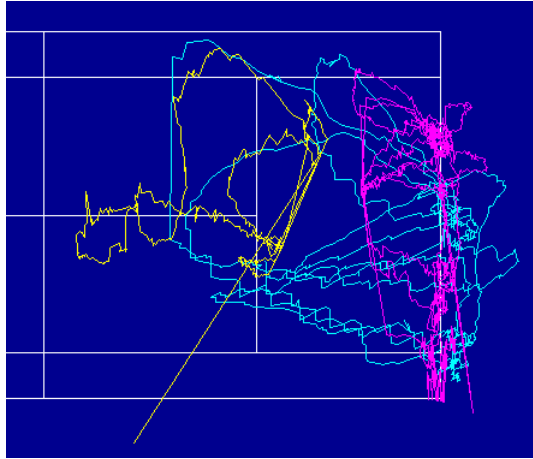


Figure 4-2: Image of the resulting trajectories from each camera

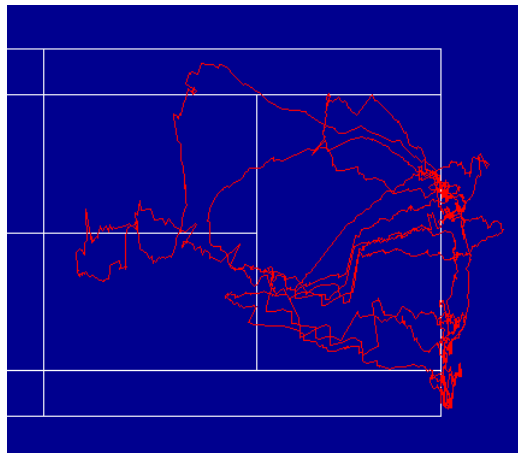
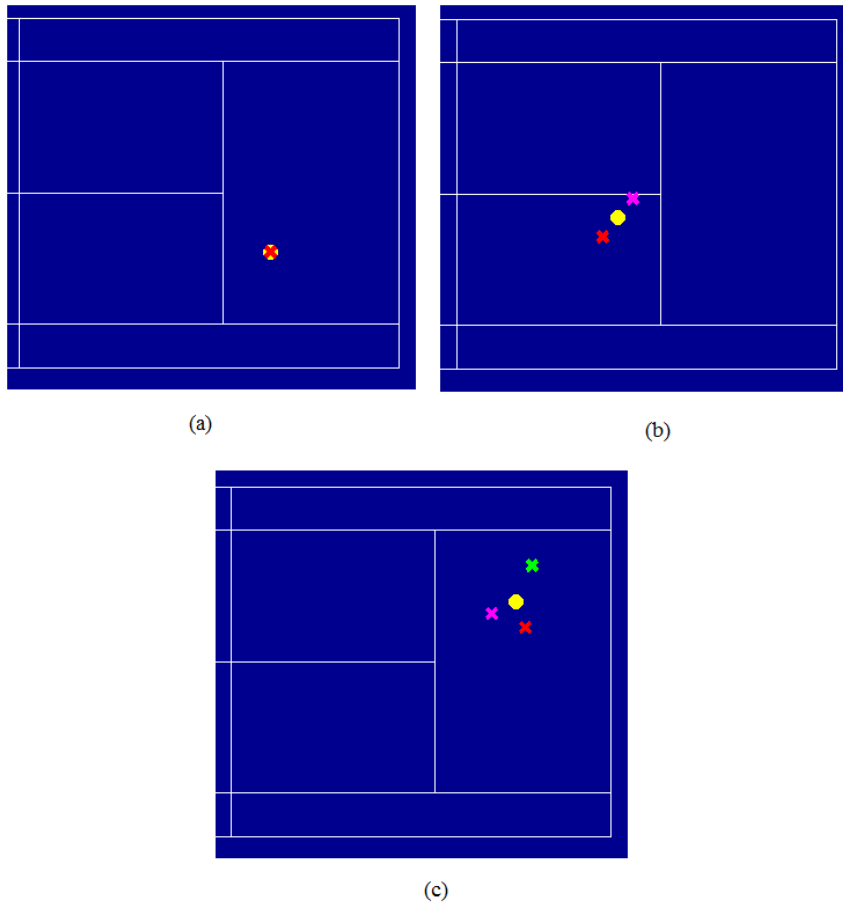


Figure 4-3: Image of the resulting trajectory from fusion

Some examples of the different possible number of cameras tracking simultaneously the player are shown in figure 4-4. In this figure, the resulting point of the fusion is shown as a yellow circle, and the resulting point of each camera tracking is shown as a cross (red, green or magenta, depending on the tracking camera).



**Figure 4-4: Examples of the possible number of cameras tracking simultaneously the player:
 (a) One camera, (b) Two Cameras, (c) Three Cameras**

Due to the location of each camera and depending on the point at which the player is on the field and on the number of cameras with the player located, the error will be larger or smaller.

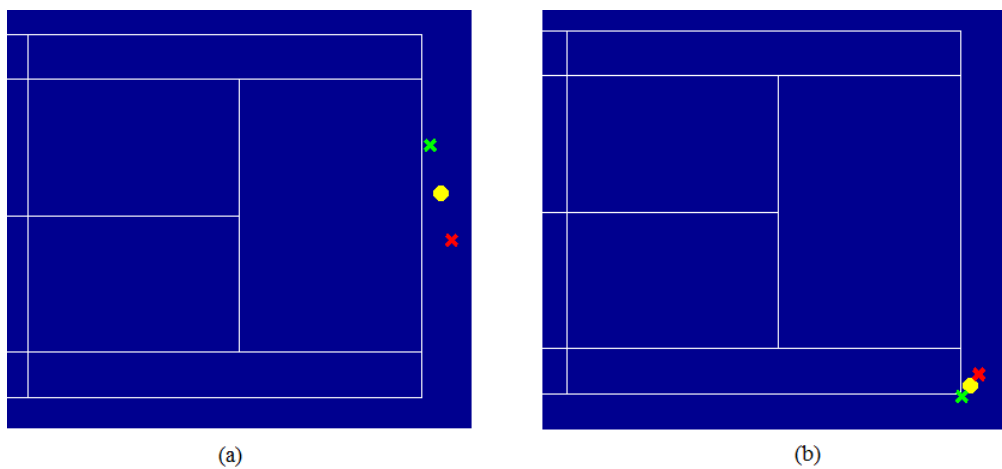


Figure 4-5: Examples of the major or minor error depending on the position of the player

In figure 4-5 two different points are shown with two simultaneously cameras tracking. In case (b), points tracked from each camera are closer between them, and in case (a), points are more distant.

Part of the position errors are caused by distortion of camera lenses. One possibility would be to compensate this distortion, as proposed in the corresponding section as a future work. In the presented case this problem is solved by the fusion method, as shown in the results. Because no ground truth is available, an objective measure of the error committed can not be calculated.

4.5 Discussion

For individual sports an acceptable result is achieved because it is the simplest case of sports.

The cameras are not positioned symmetrically and thus errors occur. As shown in Figure 4-4, depending on the region in which the player is placed, the error between the cameras is higher or lower. Furthermore, when the number of simultaneous cameras that follow the player from one frame to another changes, there is a sudden change of coordinates, particularly if the projection of the cameras is not too tight to each other.

In the analyzed videos, the tennis dataset cameras are located at a lower height than the cameras of soccer dataset (as seen in the next sections), which is reflected in more distant pairs of trajectories in the case of tennis, even though the football field is larger, which might suggest that major errors should occur.

5 Team sports

5.1 Introduction

In this section, the presented system is centered on team sports. We refer team sports for sports where two or more players move in the same areas of the field. Referees may also be moving in the same areas and between the players.

In this case, fusion is harder because all the tracked blobs do not belong to the same player as in the case of individual sports. During tracking, occlusions may occur, causing problems in tracking as losing players or generating fused blobs from more than one player.

Some examples of this kind of sports are football, basketball or volleyball.

5.2 System

In this section the global system is presented. The system block diagram is depicted in figure 5-1.

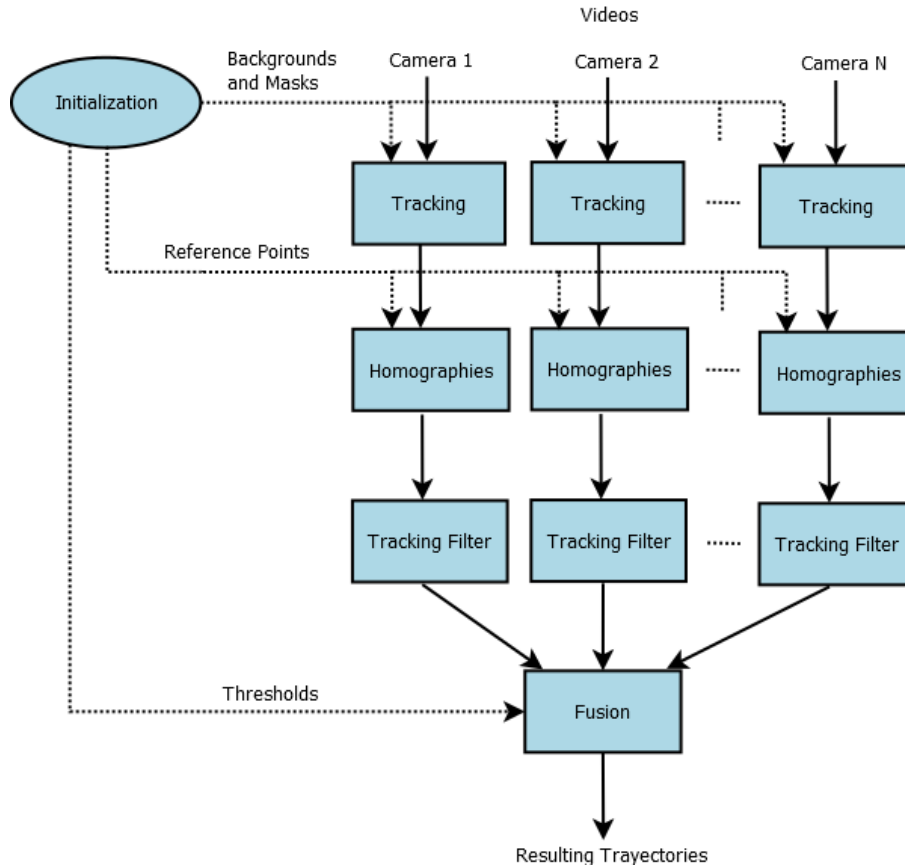


Figure 5-1: System block diagram for team sports

The system's block diagram is similar to the block diagram of individual sports but with some modifications.

The Initialization block is executed first. It generates the field representation, the backgrounds, the masks, the homography reference points and the thresholds for the fusion. The generation method of each of these elements is detailed in the corresponding section and the thresholds for fusion are defined in the following section.

The **Tracking blocks** receive the videos and the backgrounds and masks of the different cameras which are recording the players and generate the tracking data for the different perspectives.

Each **Homographies block** receives the tracking of the players from each camera and the reference points, and it generates the top view tracking from each camera.

The **Tracking filter block** receives and filters the tracking for the top view tracking from each camera. Some examples of filtering criteria are: possible regions in the field, minimum number of frames for the blob, maximum width or height for the blob. The block generates the top view filtered tracking from each camera.

Finally, the **Fusion block** receives the thresholds, which are defined in the initialization block, joins the processed tracking from each camera and generates the overall trajectories along the entire field. The fusion method will be described in the next section.

5.3 Fusion

For team sports, the fusion process is harder than for individual sports. In this case all the blobs do not belong to the same player.

For the fusion process, the following information is obtained for each blob: number of frames, initial frame, final frame, number of frames and coordinates (x and y) for each frame.

A threshold is defined to decide which blobs should be fused. All the scores under the threshold indicate that both blobs belong to the same player. The three methods to calculate the score are explained in the following subsections. Each score is calculated from two blobs, one from each tracking¹⁶. Each one of the two blobs belongs to a different tracking. The two trackings must have an overlapping field zone for the fusion. This

¹⁶ In this case the tracking is defined as the set of blobs that is fused with another set of blobs

threshold can be calculated from a training sequence with the evaluation system presented in section 5.4.

When a score is obtained for each pair of blobs from two different trackings, the next step is to obtain a list of blobs associations (LOA) containing one row for each blob of one of the trackings. Each row indicates which blobs in the other tracking should be fused with the blob that corresponds to that row. A similar LOA for the other tracking is not necessary to calculate because it is redundant, although sometimes it is useful for simplifying further processing (e.g. in the step where the resulting blobs of the fusion process are calculated). Figure 5-2 shows an example a LOA.

	1	2
1	5	15
2	20	0
3	0	0
4	21	0
5	4	19
6	0	0

Figure 5-2: Example of LOA

The information extracted from the figure is: blob 1 of the first tracking is fused with blobs 5 and 15 of the second tracking (if there were more blobs fused, more columns will appear). Blob 2 of the first tracking is fused with the blob number 20 of the second tracking. Blob 4 of the first tracking is fused with the blob number 21 of the second tracking. Blob 5 of the first tracking is fused with blobs 4 and 19 of the second tracking. Blobs 3 and 6 do not fuse with any blob.

Finally, for the fusion, the LOA is processed to get the group of blobs in each camera that should be fused.

The three proposed fusions have been developed through an incremental development. Each improvement is proposed to solve problems found in the previous fusion. In the tests (see section 5.5), the results of the three fusions are shown to compare the performance of each fusion.

In the following three subsections each of the three fusions developed are described. The first of the fusions uses only information of the position of each player. The second fusion adds to the first fusion color information of the different uniforms to discriminate between players. The third fusion improves the previous two fusions, adjusting spatially the projection of the trajectories of each player.

5.3.1 Basic fusion

Given two blobs, the score is defined as the mean square distance per frame, for those frames in which there are both blobs.

$$\text{Score}(\text{blobI}, \text{blobJ}) = \begin{cases} \sum_{f=n}^m ((x_{f,i} - x_{f,j})^2 + (y_{f,i} - y_{f,j})^2), & \text{If there are frames in which both blobs appear} \\ \infty, & \text{Else} \end{cases}$$

Where $x_{f,i}$ and $y_{f,i}$ are the coordinates of the blob I in the frame f , $x_{f,j}$ and $y_{f,j}$ are the coordinates of the blob J in the frame f , and $n \dots m$ are the frames where blob I and blob J are tracked simultaneously.

5.3.2 Fusion using color information

This type of fusion is the first improvement introduced. The condition when score is not equal to infinity (“if there are frames in which both blobs appear”) is changed to: If there are frames in which both blobs appear and if both blobs wear the same uniform.

This improvement aims increasing the precision parameter, because false fusions are reduced.

5.3.3 Fusion adjusting correspondence between homographies

Sometimes, resulting trajectories from the two trackings does not fit correctly. This malfunction can be due to imperfections of the camera lens, different camera orientation and height, etc. This problem is observed in the trajectories of the left side of the figure 5-3. As detailed in Section 6, some of these corrections are proposed as possible future work. In these cases, after seeing the resulting trajectories of the homography projection (for example with a training sequence), the points that define the homography can be changed to obtain a better fit.

Another solution that allows more freedom is to apply a correction to the trajectories after applying the homography. In this way additional problems can be corrected, compared to

those obtained simply by changing the points of the homography. Some examples of these additional problems are barrel distortion and pincushion distortion.

Figure 5-3 and 5-4 show how this improvement results in a better fit between trajectories of blobs from the different trackings.

5.4 Evaluation system

For the evaluation system, in addition to the data described in the fusion section, the unique ID of the blob is needed, corresponding to the player or referee. This ID is obtained from the ground truth¹⁷. This ID allows knowing if two blobs belong to the same player and evaluating how many fusions are correct or erroneous.

Two blobs should be fused if they belong to the same player (both have the same unique ID) and if there are frames in which both blobs appear. For the fusion of two trackings, a list like the LOA described in section 5.3 (see Fig 5-2) is obtained but with the certainty that in this case the ideal list of blobs associations (iLOA¹⁸) is obtained, with all the correct fusions. This iLOA allows calculating the Precision and Recall (defined at the end of this section) when it is compared with the experimental list of blobs associations (eLOA).

After obtaining the iLOA and defining a set of values for the threshold defined in Section 5.3, eLOAs are obtained. Each eLOA is compared to the iLOA, obtaining the successful and wrong fusions in each case.

If a blob number is found in the same line in both lists, the fusion is correct. The total number of elements in the iLOA corresponds to the total number of expected fusions. The total number of elements in the eLOA corresponds to the total number of fusions obtained by the system.

After obtaining the necessary data, the values of recall and precision for the fusion process are calculated. Definitions of recall and precision are extracted from [13]: Recall is the fraction of accurate associations to the true number of associations; Precision is the fraction of accurate associations to the total number of achieved associations. Let ξ_{Ω} be the

¹⁷ In the ISSIA soccer dataset it is extracted from the ground truth tracking of each of the 6 cameras. The blob ID for the blobs of a player is the same for the six tracking files.

¹⁸ The LOA, iLOA and eLOA are “classes”, which are instantiated as specific lists (e.g., FCeLOA, RiLOA) during the different fusions performed (see section 5.5.2).

ground truth for pairs of trajectories on the overlapping region and let E_Ω be the estimated results. Then, R and P are calculated as:

$$Recall = \frac{|\xi_\Omega \cap E_\Omega|}{|E_\Omega|}$$

$$Precision = \frac{|\xi_\Omega \cap E_\Omega|}{|\xi_\Omega|}$$

5.5 Adjustments, testing and results¹⁹

For the testing, the used videos are from the ISSIA Soccer Dataset (see section 2.3.1). The available ground truth tracking consists of an ideal tracking of each camera in which, besides the position and size of each player in each frame, the unique ID of each blob is provided. It allows knowing which player corresponds to each blob. This tracking had some annotation errors that were corrected when they were detected²⁰.

There are some important definitions that facilitate understanding the rest of the section:

- Facing cameras: Overlapping cameras covering an area of the field. There are three pairs of facing cameras: camera 1 and camera 2, camera 3 and camera 4, camera 5 and camera 6.
- Regions: The resulting fusion of each pair of facing cameras is named region. There are three regions, one for each pair of facing cameras.
- Field: The result of combining the three regions, covering all the field area, by fusing the two overlapping areas of the regions (region of cameras 1-2 and region of cameras 3-4, and region of cameras 3-4 and region of cameras 5-6).

To get the fusion of the field, first facing cameras are fused and then the resulting regions are fused, generating the fusion of the field.

The number of correct fusions for each fusion ground truth is represented in the table 5-1. In the case of the base system tracking, the trajectories are more fragmented than in the ground truth tracking, causing a greater number of fusions.

Tracking	1 with 2	3 with 4	5 with 6	Total facing cameras fusions	Regions fusion
Ground truth tracking	23	75	39	137	106
Base system tracking	159	462	267	888	132

Table 5-1: Number of correct fusions

¹⁹ A web page has been created where some videos with the result of the system have been published. The graphics of Precision and Recall from annexes F, G, and H have been added with higher resolution.

<http://www-vpu.eps.uam.es/publications/DetectionAndTrackingInMulticameraSportsVideo/>

²⁰ A document with the corrections is available in the created web page.

The fusion threshold takes values between 0 and 50. The incremental approach developed for fusion is explained in section 5.5.1.

There are two different evaluations of the system used, one for each available trackings, which are described in sections 5.5.3 and 5.5.4.

5.5.1 Fusion incremental development

For each of the two available tracking data (the one provided in the ground truth of individual cameras and the resulting tracking of the system with the modifications) the result of the three types of fusion described in section 5.3 have been analyzed.

The results using basic fusion are presented with the names of GT1 and BS1, depending on the tracking used, Ground Truth (GT) or Base System (BS).

With the basic fusion and thanks to the unique ID of the blobs, to simulate the results to achieve an ideal tracking which discriminate between the different uniforms of the people in the field is possible. The results using this fusion are presented with the names of GT2 and BS2, depending on the tracking used, Ground Truth (GT) or Base System (BS).

Given the results of the first two methods, fusion between cameras 5 and 6 shows worse results than in other pairs. It is due to the homography does not fit properly, for the reasons discussed in section 5.3.3 as imperfections of the lens or different camera orientation and height. In figure 5-3, an example of the problem is shown. Two fragments of the resulting trajectories are presented from the player with unique ID = 104 of the facing cameras 1 and 2 (right) and 3 and 4 (left). The vertical distance of trajectories from cameras 5 and 6 is significantly higher than the distance between the trajectories from cameras 1 and 2.

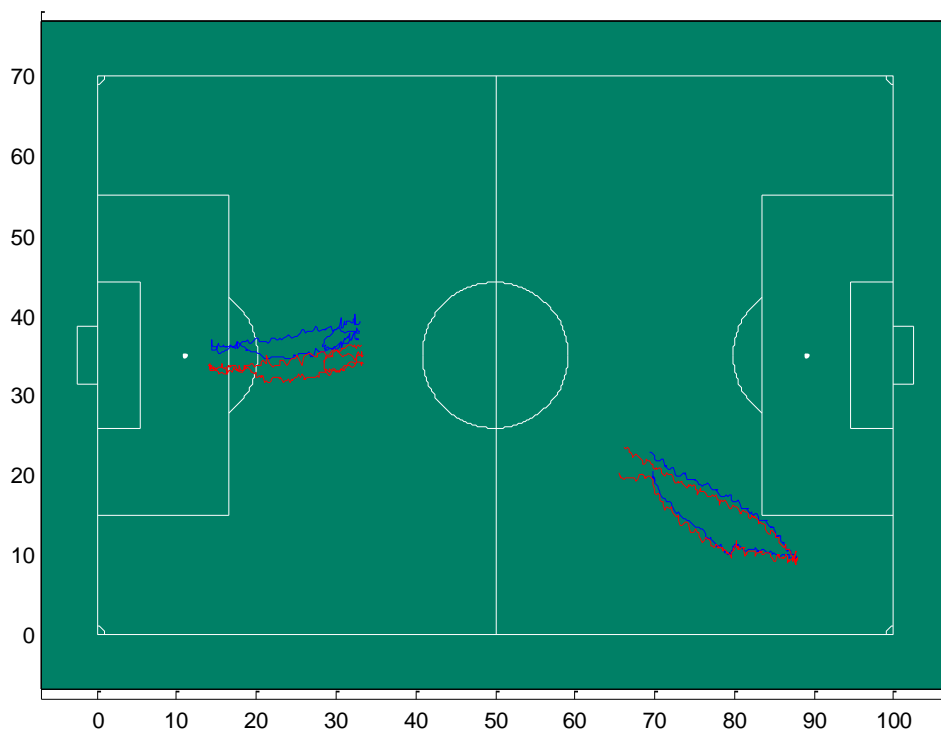


Figure 5-3: Example of trajectories before applying the homography without correction

Results named GT3 and BS3 use the same fusion than GT2 and BS2, but correcting the explained error. The example shown in Figure 5-3 with the correction of the points of the homography is presented in Figure 5-4.

The correction is done heuristically and manually after seeing the results. In a real system, the correction would be done at initialization. From a warming up video with players moving around the field, the trajectories of players are extracted and the correct fit can be obtained.

The improvement can be seen on the results of the cameras 5 and 6, presented in the annexes of results, described below. In the tests, only homographies of the facing cameras 5 and 6 are corrected. As this correction is not compensated in the other facing cameras, the results of regions fusion are slightly worse, but the goal is to show that the correction of the homography can significantly improve the results of the corrected trajectories.

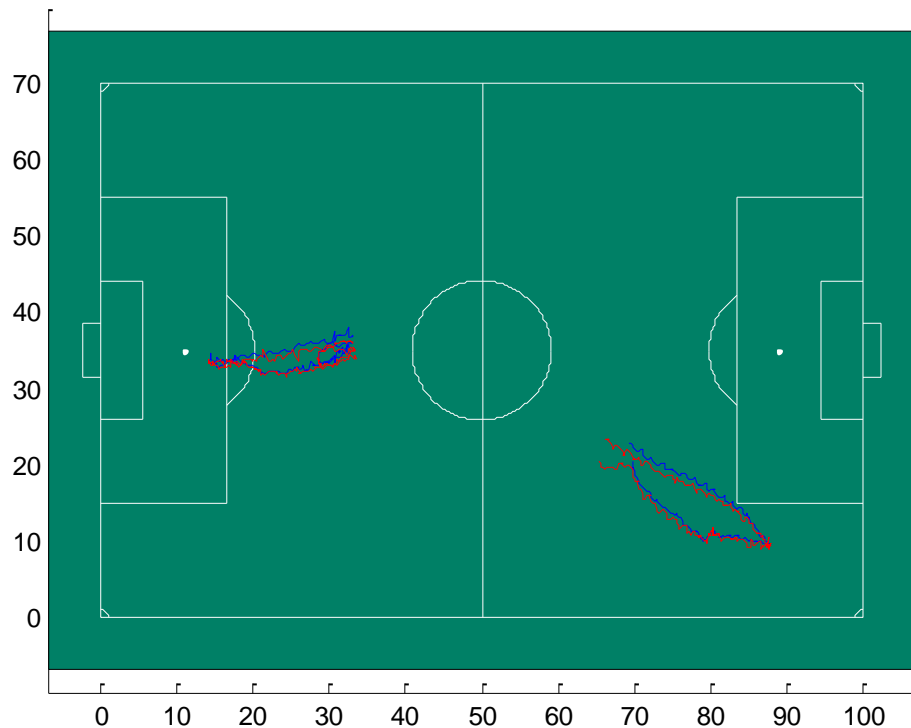


Figure 5-4: Example of trajectories after applying the homography with correction

5.5.2 Steps followed for each validation scenario

The steps for obtaining the fusion of the field, resulting in the trajectories of players, and to get the data needed (instances of LOAs) for the evaluation of the system are presented.

Steps for obtaining the trajectories of players (fusion of the field):

- 1- The annotated blobs are obtained for the different frames after the tracking. In the case of the ground truth tracking, the result is directly the ground truth tracking file, and for the base system tracking the result is the generated output file (described in section 3.3.3).
- 2- The homography corresponding to each of the camera tracking is applied to obtain the top view trajectories of each camera tracking.
- 3- The first experimental lists of blobs associations (eLOAs) are calculated between each two groups of blobs belonging to facing cameras, resulting in Facing Cameras experimental lists of blobs associations (FCeLOAs). There are 3 fusions in total resulting in 3 FCeLOAs, one for each pair of facing cameras.

- 4- The facing cameras blobs are fused according to the obtained fusion lists (FCeLOAs). After this step, the 6 trackings are reduced 3 trackings, one for each region.
- 5- The second fusion lists of blobs associations are calculated between the groups of blobs of the different overlapping regions (region of cameras 1-2 and region of cameras 3-4, and region of cameras 3-4 and region of cameras 5-6. See figure 5-7 to see an example with the resulting overlapping areas), resulting in Regions experimental lists of blobs associations (ReLOAs). There are 2 fusions in total resulting in 2 ReLOAs.
- 6- The blobs of the different regions are fused according to the calculated ReLOAs. After this step, the final field tracking with all the trajectories is obtained (see figure 5-9 to see an example of the resulting field trajectories).

Steps for obtaining the results of the evaluation system:

- 1- Making use of the FCeLOAs, the evaluation of the facing cameras is made. The ideal lists (FCiLOAs, different for each used tracking due to they depend on the resulting tracking blobs) are calculated with the unique ID (The way to get the unique ID is explained in each scenario: in the ground truth tracking is the true ID contained in the tracking files, and in the base system is obtained with spatial and temporal similarity between blobs of the base system and the ground truth) and compared with FCeLOAs (the lists obtained with the different thresholds and fusions), obtaining Precision and Recall for facing cameras.
- 2- For the evaluation, in the fourth step for obtaining the regions fusion, the fusion ground truth (ideal list) is used to prevent that the errors in this first stage affect in the evaluation of the next stage. If the ideal fusion is not applied for facing cameras fusion, the second evaluation cannot be calculated. A unique ID cannot be assigned to the resulting blob of the fusion of blobs from different players. Furthermore, when two blobs belonging to a player are not fused in the first stage but are fused in the second stage, two correct fusions are obtained instead of one, which changes the final results. Note that in the process for obtaining the trajectories of the players (without evaluation) the ground truth is not used (ideal lists, unique IDs...).
- 3- Step 1 for obtaining the results of the evaluation system is repeated, but in this case with the RiLOAs and ReLOAs. The result obtained is the precision and recall for the region fusion.

5.5.3 GT tracking based validation scenario

The first testing uses as tracking the ground truth tracking files provided in the dataset. There are six files, one for each camera, containing the tracking for each player. The blob ID for the blobs of a player is the same for the six tracking files, and is called unique ID. For example, the blobs of the player with unique ID = 5 have that identifier (ID=5) for all the tracking files. The unique blob ID is not used to facilitate the fusions. It is only used to evaluate after the fusion, which allows obtaining Precision and Recall as described in the evaluation system of the section 5.4. It is used in this way to simulate that the tracking is obtained from a real system and not from the ground truth, with the advantage that quantitative results of performance can be obtained.

The first step is to apply the homography corresponding to each of the six tracking files to obtain the top view trajectories. The result is shown in figure 5-5.

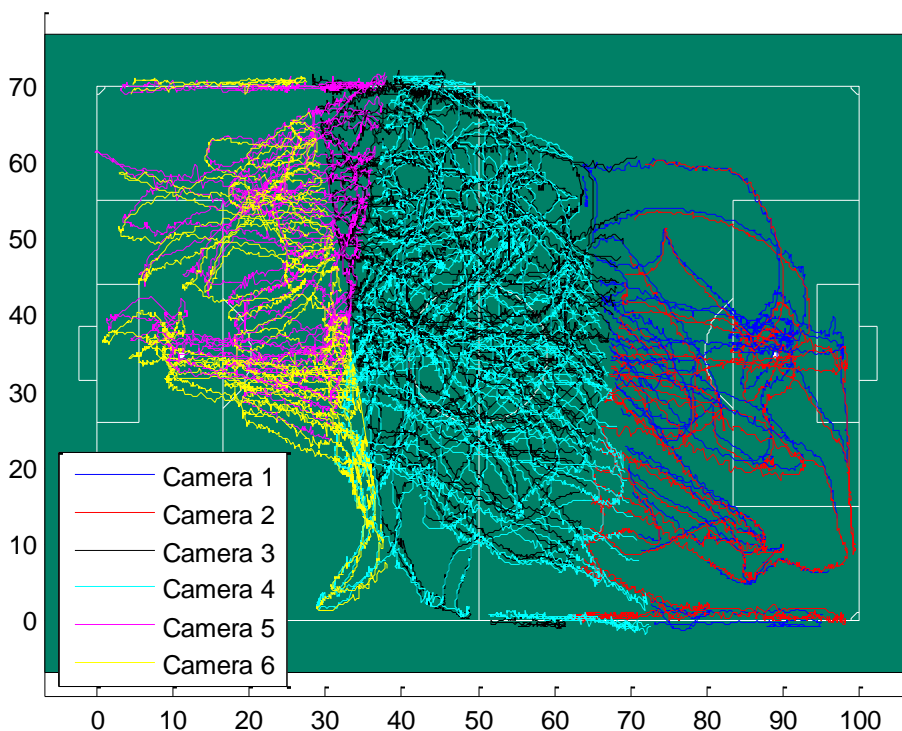


Figure 5-5: Top view tracking for each camera

Then, the facing cameras fusion is calculated. Results from fusion between facing cameras are shown in annex F, in figures F-1, F-2, F-3 and F-4. Figures F-1, F-2 and F-3 show the precision and recall for the fusion of the facing cameras 1 with 2, 3 with 4, and 5 with 6, respectively. Figure F-4 is obtained joining the results obtained from all the facing cameras

fusion. The result is not the average from the 3 pairs because the number of fusions in each pair of cameras is not the same. Below is Figure 5-6, equal to the figure F-4.

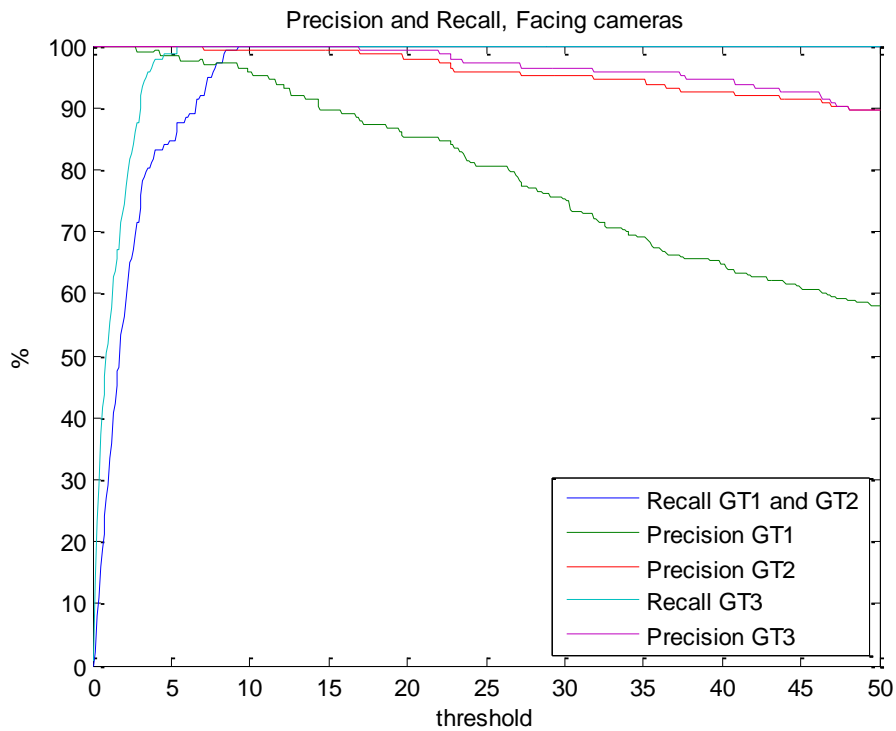


Figure 5-6: Precision and Recall from fusion of facing cameras with Ground Truth tracking

The ideal result of the facing cameras fusion is presented in figure 5-7. In the figure, each color (red green and blue) represents a region.

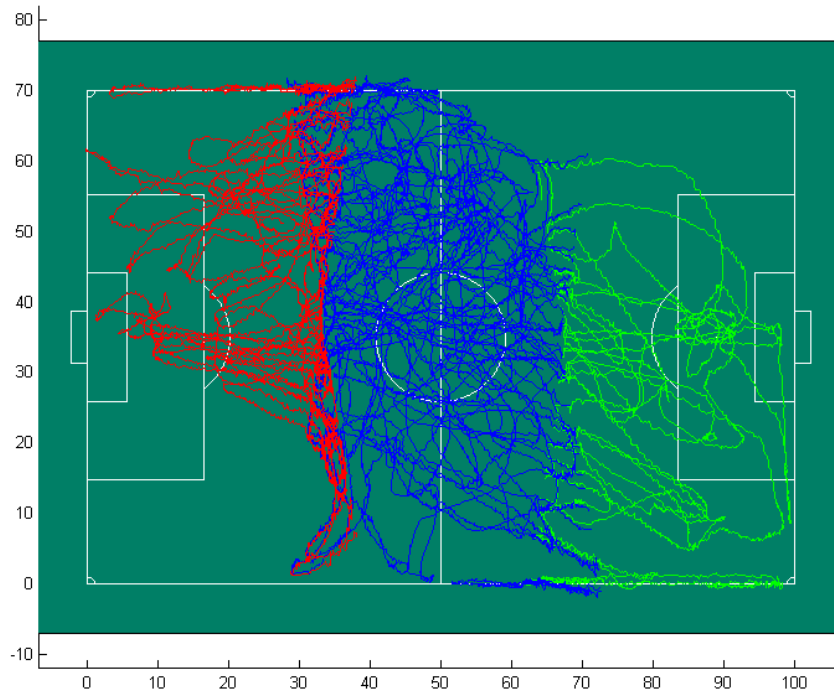


Figure 5-7: Resulting trackings from the fusion of facing cameras

Finally, the regions fusion is calculated. In annex F, figure F-5 shows the results of fusing the three regions. Below is Figure 5-8, equal to the figure F-5.

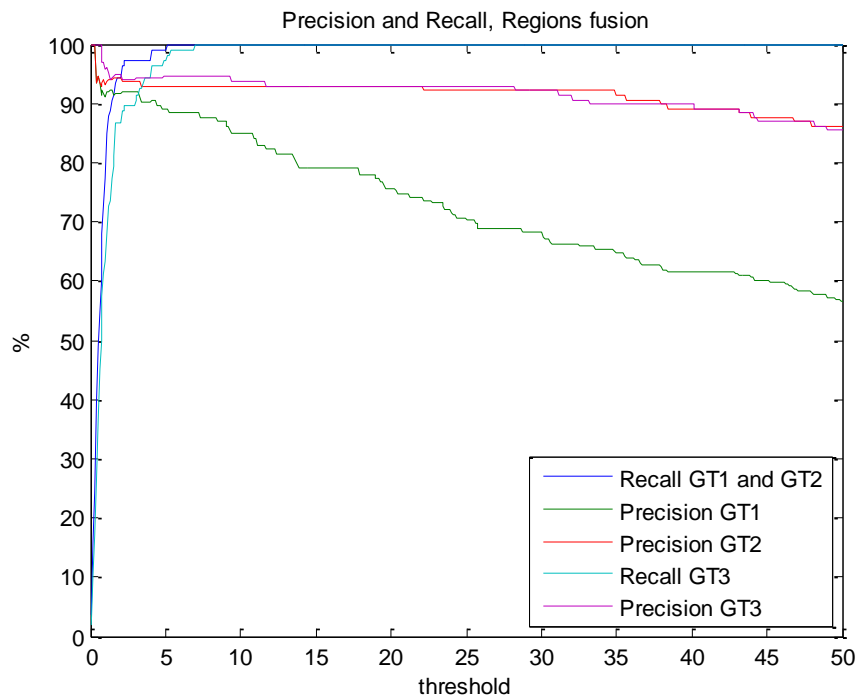


Figure 5-8: Precision and Recall from fusion of the different resulting regions with Ground Truth tracking

The ideal result of the field fusion is presented in figure 5-9. In the figure, each player is represented with a different color.

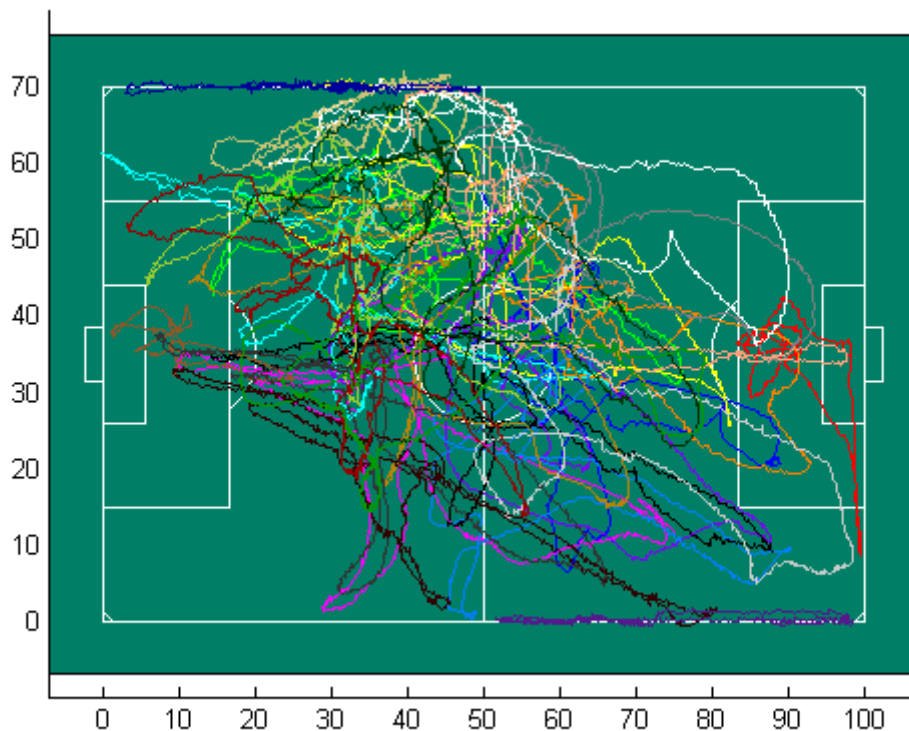


Figure 5-9: Resulting tracking from the region fusion

On annex J, the resulting trajectories from ideal and experimental fusion for the ground truth tracking scenario are presented in figures J-1 and J-2.

5.5.4 Base system tracking based validation scenario

The second testing uses the tracking module of the base system described in annex A with the modifications described in section 3.3. The average time of the tracking processing in the football videos is 7,7 FPS (lower than in the tennis videos due to the higher resolution of the football videos).

The first step is to apply the homography corresponding to each of the six ground truth tracking files and to each of the six base system tracking files to obtain the top view trajectories of both trackings.

To get the unique ID of the base system tracking blobs, the score defined in section 5.3.1 between the top view ground truth tracking blobs and the top view base system tracking blobs is calculated for each camera. For each blob of the base system tracking module, the

corresponding blob of the ground truth tracking with the lowest score indicates the unique ID. Using this method, the identifier of the closest spatially blob (in the corresponding frames) from the ground truth tracking is obtained for each of the blobs obtained from the base system tracking. To verify that the unique ID association is right, the result of the association is presented in annex E, where 25 figures are shown (22 players and 3 referees).

After obtaining the unique ID of the blobs of each camera of the base system tracking, the steps 3 to 6 of section 5.5.2 are applied: FCeLOAs are obtained, the blobs of the facing cameras are fused according to the FCeLOAs, ReLOAs are obtained and finally the blobs of the different regions are fused according to the ReLOAs, resulting in the field trajectories. As in the case of the previous scenario, the unique blob ID is used only to evaluate Precision and Recall after the fusion, as described in the evaluation system of the section 5.4.

Precision and recall for this validation scenario are shown in annex G, which has the same structure than annex F and annex G . Figures G-1, G-2 and G-3 show the precision and recall for the fusion of the facing cameras 1 with 2, 3 with 4, and 5 with 6, respectively. Figure G-4 is obtained joining the results obtained from all the facing cameras fusion. Below is Figure 5-10, equal to the figure G-4.

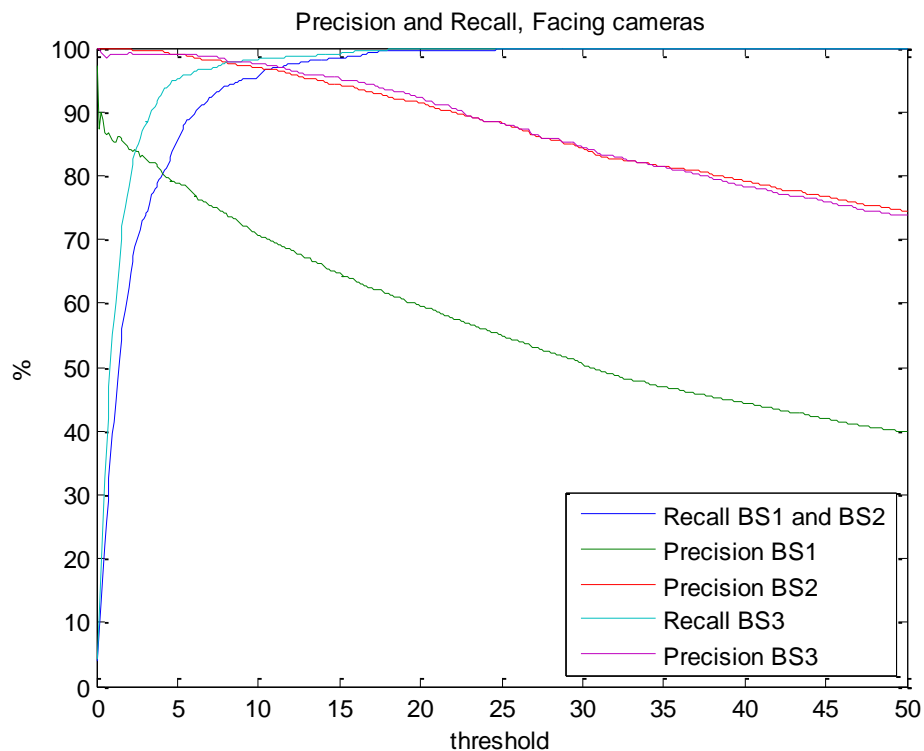


Figure 5-10: Precision and Recall from fusion of facing cameras with Base System tracking

In annex G, figure G-5 shows the results of fusing the three regions. Below is Figure 5-11, equal to the figure G-5.

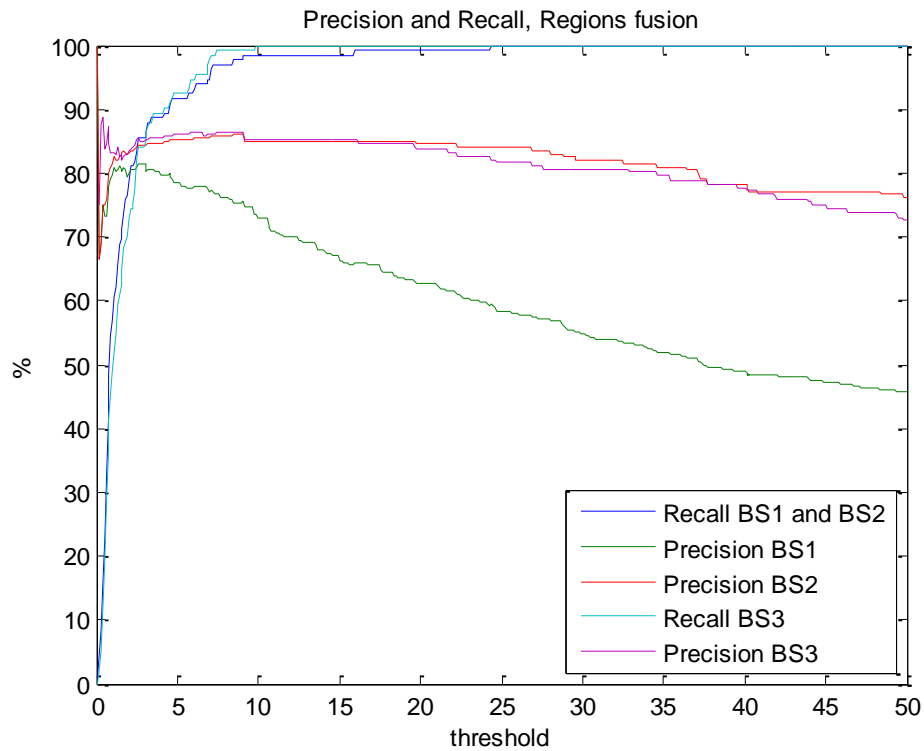


Figure 5-11: Precision and Recall from fusion of the different resulting regions with Base System tracking

On annex J, the resulting trajectories from ideal and experimental fusion for the base system tracking scenario are presented in figures J-3 and J-4.

The current system needs to be improved to be a fully automatic system (see future work in section 7.2) because it is not able to completely track the 25 players without additional post-processing (supervision or the help of the ground truth tracking). The base system tracking is not able to get the full trajectories of each player, which causes that the fusion does not link all fragments of trajectories of each player. Therefore, post-processing is required to associate multiple supervised fragments resulting from the trajectory of each player. In this case, the post processing is done under manual supervision, but as future work to design and develop a GUI module that does support this task (as in the case of initialization) is proposed, as well as work on tools that automatically improve the tracking and fusion results in order to minimize the supervision (it should be noted, that all current commercial systems use some kind of supervision). In the case of the referee, the linesmen and the goalkeepers, automating the process based on color and distance should not be complicated.

On annex K, the supervised associated fragments of the trajectories of each player are shown. The trajectories have been associated with help of the unique tracking ID of the ground truth, but in real systems ground truth is not available and therefore the trajectories must be associated with supervision. Next to each image, the complete trajectory of each player, extracted from the ground truth tracking, is shown to compare both results. Table 5-2 shows the number of resulting trajectory fragments which compose the complete trajectory.

Unique ID	200	201	202	1	105	5	6	8	9	10	11	14	20
Number of fragments	26	38	30	4	3	19	5	13	21	25	12	7	4

Unique ID	21	26	104	107	108	113	114	120	125	126	128	155
Number of fragments	11	13	13	18	14	17	12	14	18	29	23	13

Table 5-2: Number of trajectory fragments of each player

5.5.5 GT tracking results vs. base system tracking results

Annex H presents a comparative between the results obtained with each of the two types of tracking, and has the same structure than annex F and annex G. From these figures some conclusions are extracted, presented in the section 0.

6 Statistics from system results²¹

After obtaining the results for both types of systems (individual sports and team sports), some additional functionalities have been implemented to show some players statistics.

A zigzag effect occurs between two consecutive frames, as shown in figure 6-1. There are multiple sources of error that cause this problem: camera lenses, homographies, annotation, segmentation, tracking or fusion. This effect causes an error in the obtained statistics because they are higher than the real statistics. To reduce this error, the statistics are calculated every 25 frames, which corresponds to one second of video.

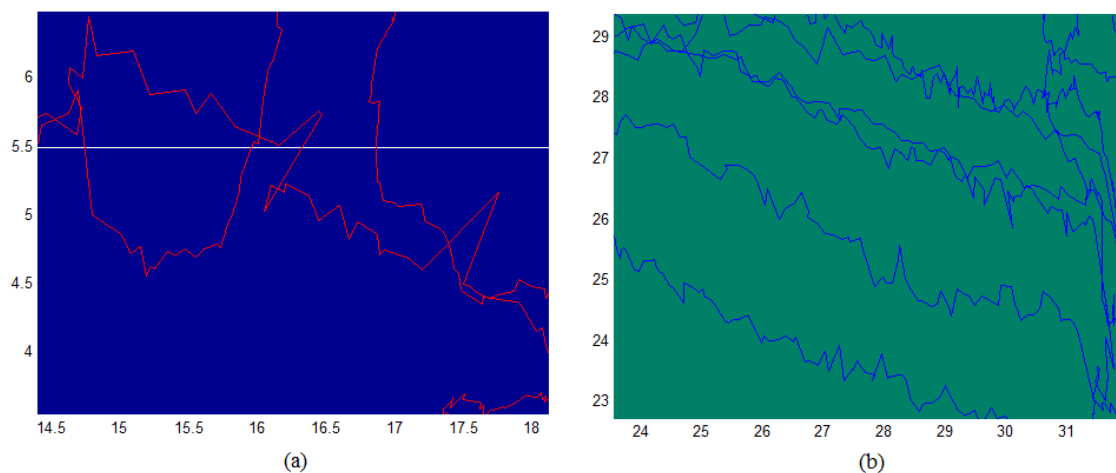


Figure 6-1: Examples of the zigzag effect in the trajectories of the tennis system (a) and in the trajectories of the football system (b)

The statistics presented are: position on the X axis, position on the Y axis, covered distance, average speed and instant speed. The position on the X axis and the position on the Y axis are directly the coordinates of the player in a frame. The covered distance is calculated by adding the covered distance in each temporal interval (a temporal interval of one second, as indicated previously). The average speed is calculated by dividing the distance covered by the elapsed time. The instant speed is calculated as the average speed but only considering the last temporal interval.

An example of the resulting statistics video is shown in figure 6-2 for the tennis system and in figure 6-3 for the football system (in this case the referee tracking is used to facilitate the visual tracking in the video).

²¹ A web page has been created to add videos and results, where two videos with the extracted statistics of the systems have been added.

<http://www-vpu.eps.uam.es/publications/DetectionAndTrackingInMulticameraSportsVideo/>

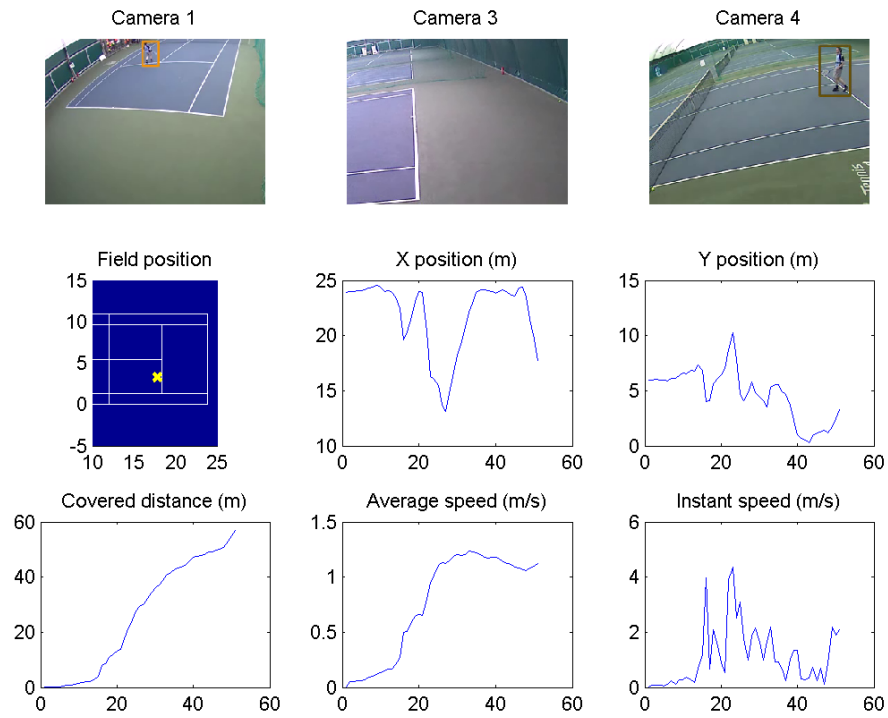


Figure 6-2: Example of the resulting tennis statistics video

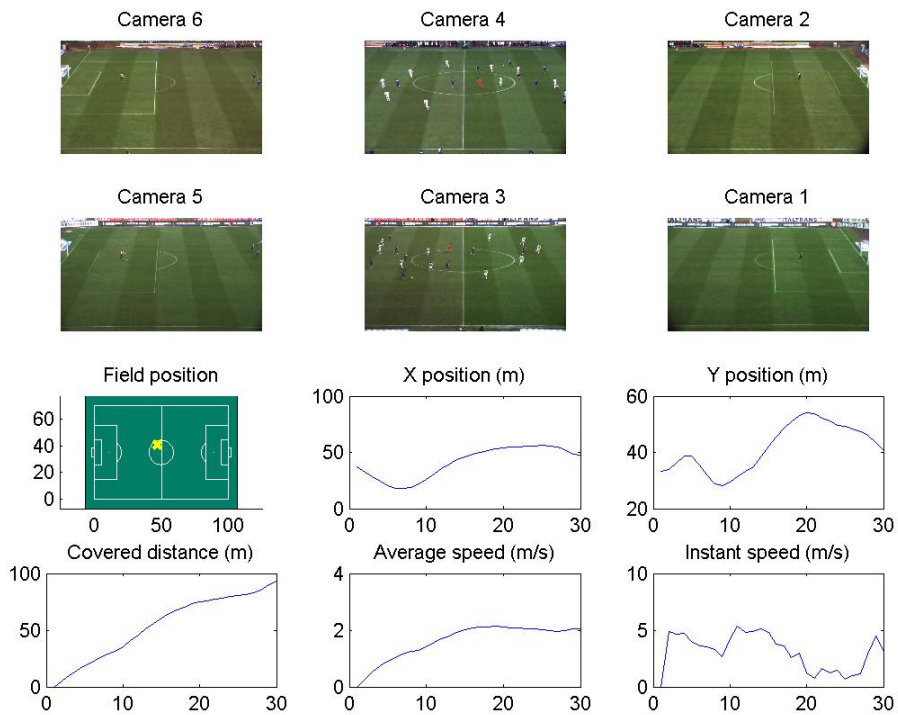


Figure 6-3: Example of the resulting football statistics video

In Annex I the extracted statistics for the tennis player and for the 25 tracked people in the football system are shown. The given statistics correspond to the full video. Getting more and more complex statistics is proposed as a future work.

7 Conclusions and Future Work

7.1 Conclusions

The main objective of this project, the design and development of a system for detecting and tracking players in a field using multicamera video, has been reached. After a previous configuration and with some supervision (for team sports), the system is able to detect and track each player in the field, and to provide some statistics. The system is complete, general and modular, easing future work improvements and modifications.

Sport videos with fixed cameras have significant characteristics that provide advantages for analyzing them with respect to general videos of video surveillance. Backgrounds are generally static and uniform, except for certain areas such as public or dynamic advertising, which can be modeled relatively simple. People tracked have specific and distinct uniform, at least between the different types of people (players from each team, goalkeepers, referees ...). This last case has a disadvantage, because the players on a team have exactly the same appearance as they wear the same uniform, which may complicate the tracking when occlusions occur.

The location of each camera is important. The ideal case is when the cameras are faced or symmetrically placed, since in these cases tracking errors are reduced in fusion process.

Better results are shown in cases with greater overlapping. This feature is observed in the case of team sports, when comparing the results of the facing cameras with the results from the fusion of regions.

Placing the cameras at higher elevations is interesting because it reduces the tracking error. The greater height of the camera, the smaller area of pixels is projected in the plane of the field and therefore ensures greater precision to the homography projection, but if the camera is located too high, the identification of the player uniform can be difficult. Also, using the base mid-point reduces error because, as discussed in the corresponding section, it reduces the distance between the center of the player base and the point chosen.

In the case of individual sports the system is relatively simple. If the background keeps static and the tracking is working relatively well, the result is appropriate and complies with the objectives. As there are no occlusions, it greatly facilitates the tracking and determination of who owns each blob.

In the case of team sports, the main problems are occlusions and regions with low or without overlapping. These systems are on which most improvement is needed, as seen in

state of art and in experimental results. There are many methods of fusion and many parameters which, when combined properly, may contribute to better results. In the upgrades made, with relatively simple changes, significant improvements are obtained in the results. As shown in the experiments, in the case of the real tracking system a supervised post-processing for obtaining the complete paths is required, whose replacement by an automatic process is proposed as future work.

7.2 Future Work

Some future work lines are:

- **Background extraction:** The extractor used is relatively simple. Work could focus on making an extractor more complex or one with lower computational cost, for example selecting invariant pixels (or within margins) for a certain number of frames.
- **Graphical User Interface:** As described in the different sections of the system initialization, individual modules are configured manually in the corresponding script. Also for the supervised fusion of partial trajectories (see section 5.5.365), the tasks is currently done manually. A real-time interactive GUI should be developed for the system operators or system supervisors.
- **Tracking system:** This perhaps is the line with more possibilities for development. The system used was slightly adapted from one designed for video surveillance. This system is a good base but there are many improvements and changes that could produce better results. Some proposals are:
 - **Adaptative background:** Taking advantage of the specific characteristics of the background in each sport (public, dynamic advertising, ...)
 - **Detect colors of the uniform of the players:** With this improvement it could be applied the fusion improvement presented and simulated in the team sports fusion section.
 - **Real time:** Designing a system that allows tracking players and fusing trajectories in real time.
- **Individual Sports:** System presented for individual sports is relatively simple because the difficulty is less compared to team sports, but some lines of work can be developed to improve results. Applying weights to the contribution of each

camera, design a positioning of the cameras to improve the results or correcting the distortion introduced by the lens of the cameras, are some possible lines of work.

- Team Sports: As mentioned previously, there are many parameters and methods for the fusion of the trajectories in team sports videos. Future work may consist of combining some existing methods or add new ones. The post processing block to automatically connect fragmented resulting trajectories of the players in the team sports system is another line of future work.
- Performance and Tactic analysis: From the complete system, additional statistics and information can be calculated. Areas of the field where the team stays longer, placement of players on the field, area of influence of each player, etc. Also the effect of zigzag can be studied and corrected.

References

- [1] Y. Xinguo, D. Farin, **Current and emerging topics in sports video processing**. In Proc. of ICME 2005.
- [2] P. Figueroa, N. Leite, R. Barros, I. Cohen, G. Medioni, **Tracking soccer players using the graph representation**. In Proc. of ICPR 2004.
- [3] Y. Huang, J. Llach, S. Bhagavathy, **Players and ball detection in soccer videos based on color segmentation and shape analysis**. In Proc. of ICMCAM 2007.
- [4] C. Poppe, S.D. Bruyne, S. Verstockt, R.V. de Walle, **Multi-camera analysis of soccer sequences**. In Proc. of AVSS 2010.
- [5] M. Xu, J. Orwell, G. Jones, **Tracking football players with multiple cameras**. In Proc. of ICIP 2004.
- [6] G. Kayumbi, P.L. Mazzeo, P. Spagnolo, M. Taj, A. Cavallaro, **Distributed visual sensing for virtual top-view trajectory generation in football videos**. In Proc. of CIVR 2008.
- [7] S. Choi, et al., **Where are the ball and players? Soccer game analysis with colorbased tracking and image mosaic**. In Proc. of ICAP 1997.
- [8] Liang, D., et al., **A scheme for ball detection and tracking in broadcast soccer video**. In Proc. of PCM 2005.
- [9] Tong, X., et al., **An Effective and Fast Soccer Ball Detection and Tracking Method**. In Proc of ICPR 2004.
- [10] T. Bebie, H. Bieri, **SoccerMan: reconstructing soccer games from video sequences**, In Proc. of ICPR 1998.
- [11] M. Taj, A. Cavallaro, **Multi-camera track-before-detect**. In Proc. of ICDSC 2009.
- [12] G. Kayumbi, N. Anjum, A. Cavallaro, **Global trajectory reconstruction from distributed visual sensors**, In Proc. of ICDSC 2008
- [13] N. Anjum, A. Cavallaro, **Trajectory association and fusion across partially overlapping cameras**. In Proc. of AVSS 2009.
- [14] Y.A. Sheikh, M. Shah, **Trajectory association across multiple airborne cameras**, Transactions on Pattern Analysis and Machine Intelligence, 30(2):361-367, Feb. 2008.
- [15] J. Kang, I. Cohen, G. Medioni, **Tracking people in crowded scenes across multiple cameras**. In Proc. of ACCV 2004.
- [16] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, J.-L. Van Gool, **Color-based object tracking in multi-camera environments**. In Proc. of DAGM 2003.
- [17] I.N. Junejo, H. Foroosh, **Trajectory rectification and path modeling for video surveillance**. In Proc. of ICCV 2007.
- [18] I. Sachiko, S. Hideo, **Parallel tracking of all soccer players by integrating detected positions in multiple view images**. In Proc. of ICPR 2004.
- [19] Y. L. de Meneses, P. Roduit, F. Luisier, J. Jacot, **Trajectory analysis for sport and video surveillance**, Electronic Letters on Computer Vision and Image Analysis, 5(3):148-156, Mar. 2005.
- [20] R. Hartley , A. Zisserman, **A Multiple View Geometry in Computer Vision**, Cambridge University Press, 2003.
- [21] P. J. Figueroa, N. J. Leite, R. M. Barros, **Tracking soccer players aiming their kinematical motion analysis**, Trans. on Computer Vision and Image Understanding, 101(2):122-135, Feb. 2006.
- [22] T. Misu, S. Gohshi, Y. Izumi, Y. Fujita, M. Naemura, **Robust tracking of athletes using multiple features of multiple views**. In Proc. of WSCG 2004.

- [23] W. Du, J.B. Hayet, J. Piater, J. Verly, **Collaborative multi-camera tracking of athletes in team sports**. In Proc. of CVBASE 2006
- [24] Luis Caro, **Contribuciones a la detección de objetos robados y abandonados en secuencias de video-seguridad**, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Julio 2011.
- [25] Juan C. SanMiguel, José M. Martínez, **A semantic-based probabilistic approach for real-time video event recognition**, Computer Vision and Image Understanding, 116(9): 937-952, Sept. 2012.
- [26] Javier García Ocón, **Autocalibración y sincronización de múltiples cámaras PTZ**.

Annexes

A. Base System²²

A.1 Introduction

The work presented in this document starts from a video analysis system for abandoned and stolen object detection provided by the Video Processing and Understanding Lab.

This system is designed to work as part of a video-surveillance framework capable of triggering alarms for detected events in real time. This requirement imposes limits on the time complexity of the algorithms used in each of the analysis modules. After the initial frame acquisition stage, a foreground mask is generated for each incoming frame at the Foreground Segmentation Module. This foreground mask consists on a binary image that identifies the pixels that belong to moving or stationary blobs. Then, post-processing techniques are applied to this foreground mask in order to remove noisy artifacts and shadows. After that, the Blob Extraction Module determines the connected components of the foreground mask. In the following stage, Blob Tracking Module tries to associate a unique ID for each extracted blob across the frame sequence.

A.2 Foreground segmentation module

The purpose of the Foreground Segmentation Module is the generation of binary masks that represent whether pixels belong to the background or foreground (moving or stationary blobs).

Based on the BackGround Subtraction (BGS) segmentation technique, a background model is created and then updated with the incoming frames. This initial mask then undergoes noise and shadow removal operations in order to obtain the final foreground mask for the current frame and perform connected component analysis for blob extraction. Figure A-1 depicts the block diagram of the Foreground Segmentation Module.

²² This annex has been extracted from [24], which is based on the system described in [25].

A.2.1 Background subtraction

The background model is initialized with the average value of a short sequence of training frames that do not contain foreground objects.

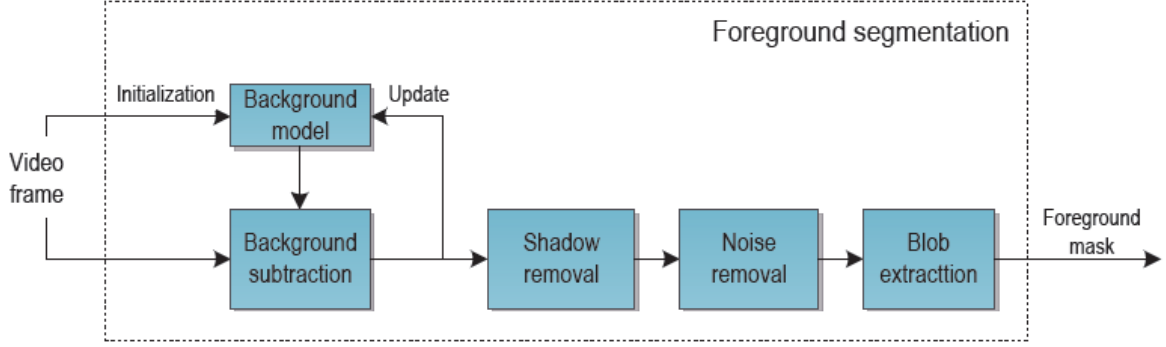


Figure A-1: Block diagram of the Foreground Segmentation Module

This model is adaptively updated to consider slow changes in global illumination conditions using a running average method. Then, the distance to the background model is calculated for each incoming frame. It consists on the squared difference between the two images (background and current), calculated around a square window for each pixel. Finally, foreground segmentation is computed by thresholding this distance according to the following equation:

$$F(I[x, y]) \leftrightarrow \sum_{i=-W}^W \sum_{j=-W}^W (|I[x + i, y + j] - B[x + i, y + j]|)^2 > \beta$$

Where W is a square window centered in each pixel, I is the current frame, B is the background model and β is the threshold for foreground segmentation.

A.2.2 Shadow removal

Shadows cast by objects and people are often misclassified as being part of the foreground due to their significant difference with the background model. Hence, high-level stages of analysis, that take as valid the data from the foreground masks (e.g., blob contour), will also be affected when adjacent shadows are wrongly considered as part of the object and therefore, their performance is decreased.

A shadow removal technique is applied to the foreground mask for removing those pixels that belong to shadows produced by moving or stationary entities (e.g., objects and people).

For this purpose, the Hue-Saturation-Value (HSV) color space is used, as it allows us to explicitly separate between chromaticity and intensity.

This approach takes advantage of the fact that for cast shadows, the change in chromaticity (hue and saturation) between the current and background image is not significant.

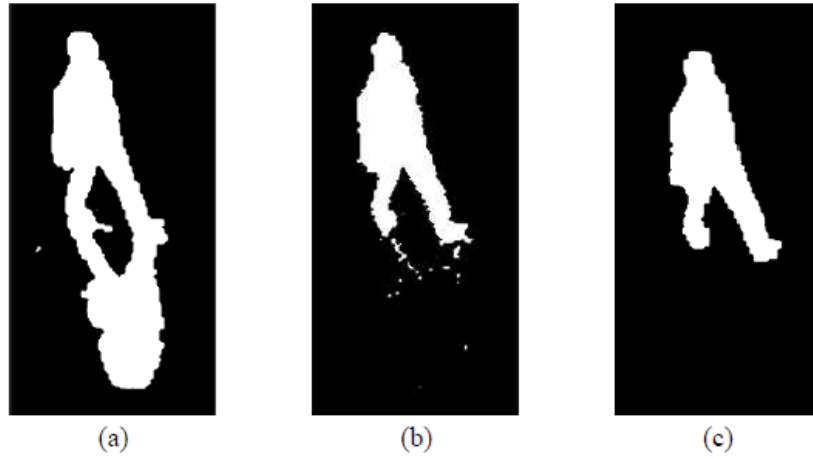


Figure A-2: Foreground mask at different stages: initial foreground mask (a), after shadow removal (b) and after noise removal (c)

The ratio intensity between both images is also computed, to detect intensity changes that are likely due to the presence of shadows. To classify a pixel as part of a shadow, the following decision function is used:

$$SP(x, y) = \begin{cases} 1, & \text{if } \alpha \leq \frac{I_v(x, y)}{B_v(x, y)} \leq \beta \wedge D_S \leq \tau_S \wedge D_H \leq \tau_H \\ 0, & \text{otherwise} \end{cases}$$

Where $SP(x; y)$ is the foreground mask that highlights pixels that belong to cast shadows at coordinates $(x; y)$; I and B are the current frame and the reference background respectively; subindexes H , S and V indicate the channel in the HSV color space; and DS and DH denote the chromatic difference between the current frame and background for both channels. The final foreground mask with removed shadows is obtained by performing logical XOR operation on SP and the mask generated by the preceding module. An example of shadow removal is shown in Figure A-2 b.

A.2.3 Noise removal

Additionally, morphological operations are performed on the resulting foreground mask for removing noisy artifacts. In particular, a combination of erosion and reconstruction operations known as "Opening by Reconstruction of Erosion" is applied. Its purpose is to remove small objects (in our case blobs due to noise), while retaining the shape and size of all other blobs in the foreground mask.

Morphological reconstruction involves an image X (the foreground mask to be processed), a marker Y and a structuring element B . In the selected approach, the marker Y is first calculated by performing the erosion operation on X , and the final mask is then obtained by performing dilation iteratively.

The size and shape of the structuring element will determine the artifacts that will be removed from the foreground mask. In Figure A-2 c we can see an example of the described operation applied to a noisy foreground mask with a 3x3 squared structuring element.

A.2.4 Blob extraction

After applying background subtraction and post-processing the obtained foreground mask, the Blob Extraction Module labels each isolated groups of pixels in the mask using Connected Component Analysis. The implemented algorithm uses 8-connectivity as the criteria to determine if pixels belong to the same connected region.

After the labeling process, some very small regions may be detected. These may be due to noise that was not correctly eliminated by the Noise Removal Module, or due to residual artifacts as a result of incorrect segmentation. In order prevent higher level modules from analyzing these regions, the ones below a certain area (in pixels) are discarded.

A.3 Blob tracking

This module performs tracking of the blobs extracted by the previous module. This is done by estimating the correspondence of blobs between consecutive frames (current and previous frames). A match-matrix MM is used to quantify the likelihood of correspondence between blobs in the previous and current frame. For each pair of blobs,

the values of this matrix are computed using the normalized Euclidean Distance between their blob centroids and their color. It is calculated as follows:

$$MM_{nm} = \sqrt{\left(\frac{\Delta X}{Xdim}\right)^2 + \left(\frac{\Delta Y}{Ydim}\right)^2} + \sqrt{\left(\frac{\Delta R}{255}\right)^2 + \left(\frac{\Delta G}{255}\right)^2 + \left(\frac{\Delta B}{255}\right)^2}$$

Where each row n corresponds to blobs in the previous frame and each column m to the number of blobs in the current frame. ΔX and ΔY are the differences in the X and Y directions between the centroids in the past and previous frame, normalized to their maximum values, $Xdim$ and $Ydim$ (the frame dimensions). ΔR , ΔG and ΔB are the differences between mean R , G and B color values, also normalized to the maximum value (255 for 8-bit RGB images).

Then, the correspondence for each blob m is defined as the index i ($i = 1 \dots n$) that presents minimum value MM_{im} .

B. 2D Projective transformations: homographies²³

2D projective geometry is the study of the properties of the projective plane P^2 which are invariant under a transformation group known as projectivities or homographies.

A homography is a bijective transformation in the projective space given by $P^2 \rightarrow P^2$ such that a straight line is transformed as a straight line.

The projectivity is defined as:

$$h(m) = m' = H * m, \quad m, m' \in P^2$$

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \rightarrow \begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

Where H is a nonsingular matrix of 3×3 . We say that m' is the linear transformation H of m . This transformation is one-one between two 2D planes, whose points are represented homogeneously by m and m' . Ie, a point on a 2D plane has a single corresponding to a point other 2D plane.

Due to the nature of the homogeneous coordinates, two proportional arrays define the same homography because points $[X, Y, Z]$ and $[kX, kY, kZ]$ are the same projective point.

Isometric transformation is a Euclidean transformation. This transformation defines a rotation followed by a translation. Isometric projective transformation is shown in figure B-1.

²³ This annex has been extracted and translated from [26].

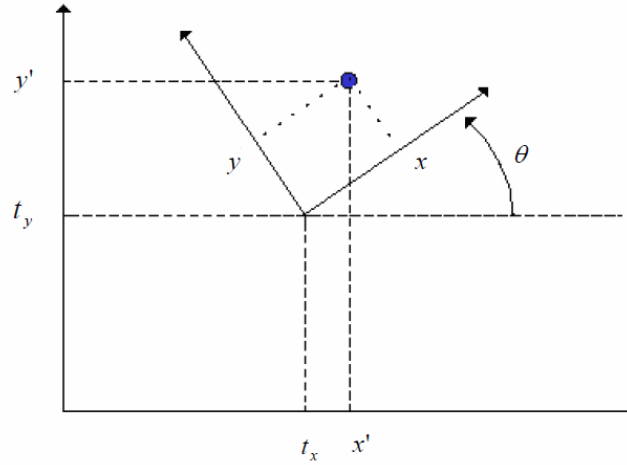


Figure B-1: Euclidean projective transformation

This figure corresponds to the transformation of euclidean coordinates $(x, y) \leftrightarrow (x', y')$, where θ is the angle of rotation of the axes and (t_x, t_y) is the offset of the origin. This transformation can be written in Cartesian coordinates as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} = R \cdot \begin{bmatrix} x \\ y \end{bmatrix} + t$$

Or in homogeneous coordinates as the projectivity:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & t_x \\ \sin(\theta) & \cos(\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

With $0^T = [0 \ 0]$. The inverse transformation $(x', y') \leftrightarrow (x, y)$ is given by:

$$\begin{bmatrix} x \\ y \end{bmatrix} = [R']^{-1} \cdot \left(\begin{bmatrix} x' \\ y' \end{bmatrix} - t' \right)$$

Since the matrix R is orthonormal (being a rotation matrix), R^{-1} is its transpose. Defining $R^T = R^{-1}$ and $t' = R^T t$, is obtained:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} R^T & t' \\ 0^T & 1 \end{bmatrix} \cdot \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}$$

C. Generated backgrounds

On this annex, the different generated backgrounds are presented.

C.1 Tennis backgrounds



Figure C-1: Background of camera 1



Figure C-2: Background of camera 3



Figure C-3: Background of camera 4



Figure C-4: Background of camera 5



Figure C-5: Background of camera 6



Figure C-6: Background of camera 8

C.2 *Football backgrounds*



Figure C-7: Background of camera 1



Figure C-8: Background of camera 2



Figure C-9: Background of camera 3



Figure C-10: Background of camera 4



Figure C-11: Background of camera 5



Figure C-12: Background of camera 6

D. Generated football masks

On this annex the different masks used for the ISSIA football dataset are presented. The original video perspective for each mask can be found in backgrounds of annex C. Black pixels correspond to '0' and white pixels (grey in document) correspond to '1'.

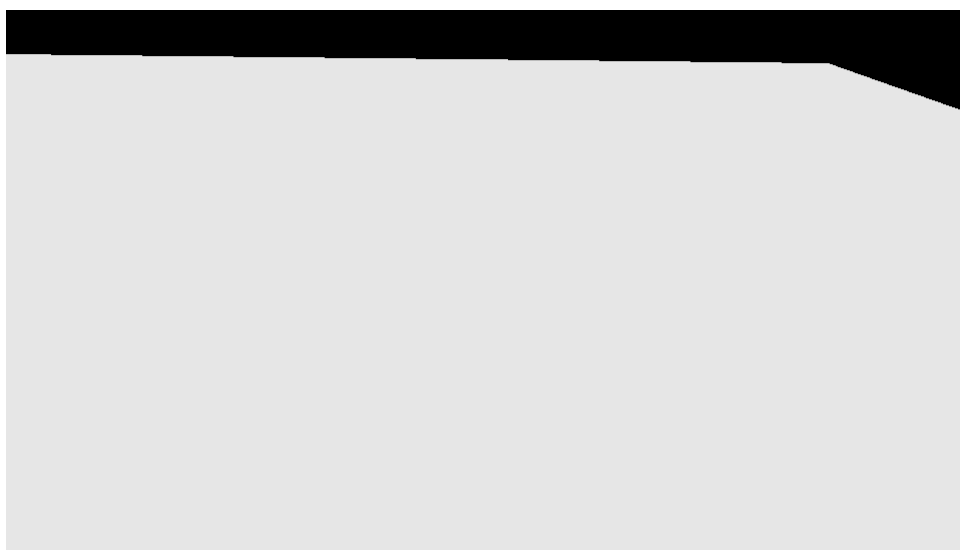


Figure D-1: Mask for camera 1

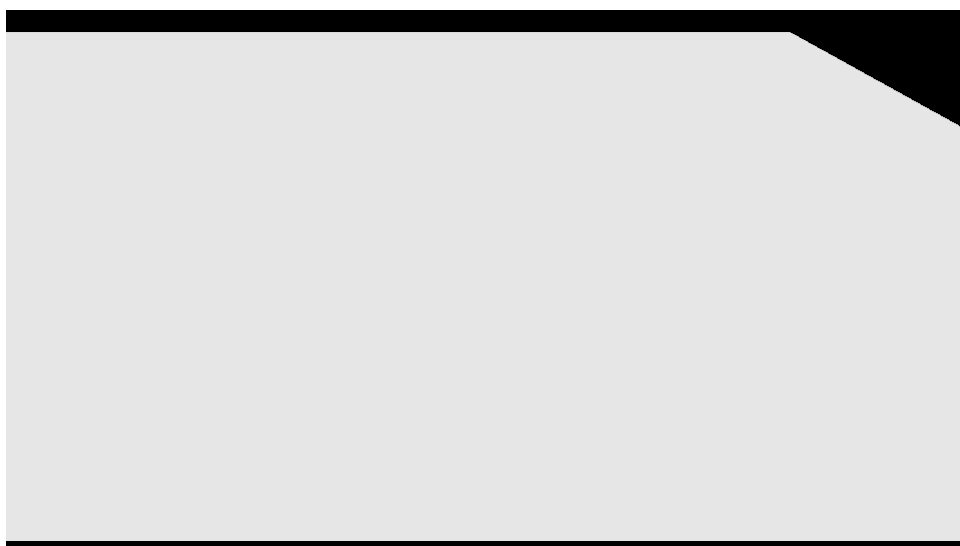


Figure D-2: Mask for camera 2

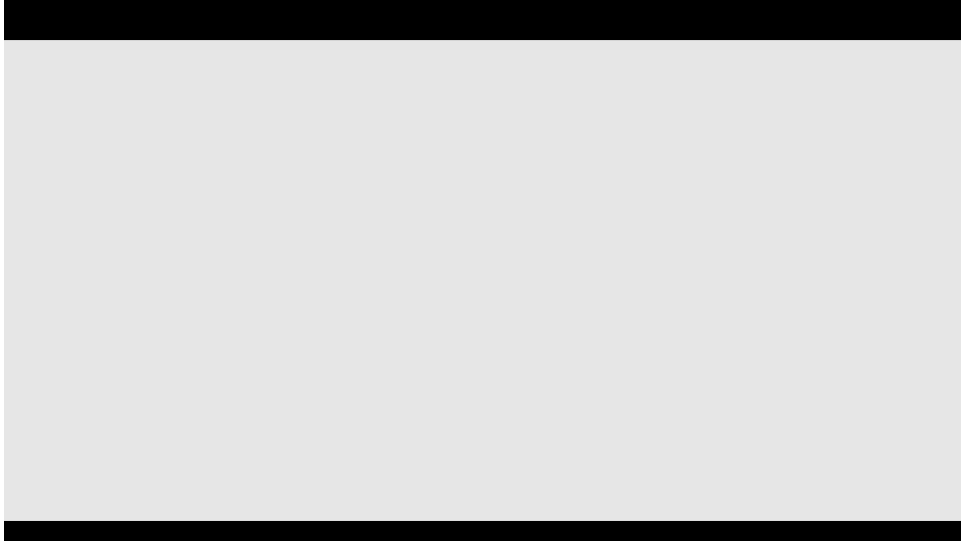


Figure D-3: Mask for camera 3

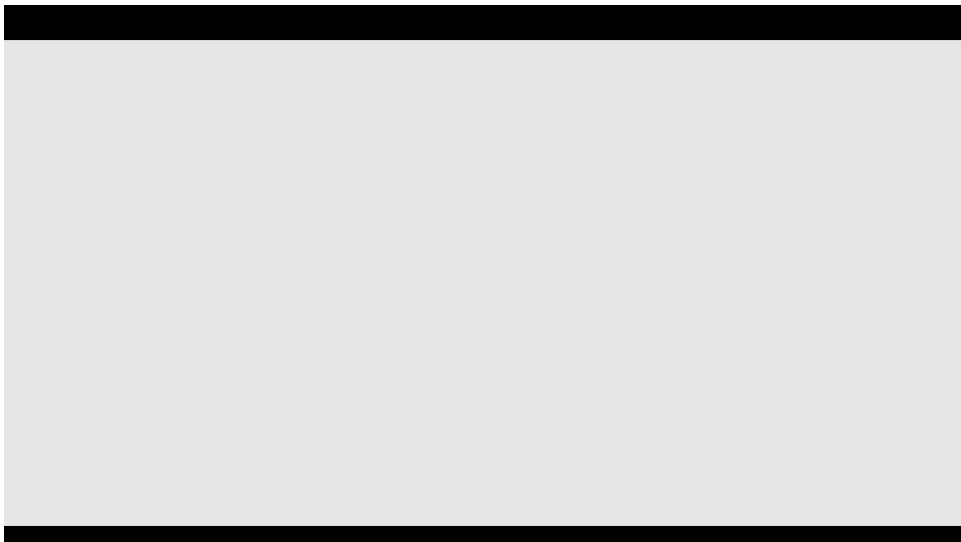


Figure D-4: Mask for camera 4



Figure D-5: Mask for camera 5



Figure D-6: Mask for camera 6

E. Representation of blobs belonging to the same player after obtaining automatically the unique ID of the base system blobs

On this annex the relation between the fused ground truth tracking and the result of the fusion of faced cameras using the base system tracking and the method described in section 5.5.4 for obtaining the unique ID of each blob is presented.

The green trajectory shows the full trajectory of each player, extracted from the ground truth tracking. The red and blue trajectories show all the blobs which have obtained the unique ID of the player with the green path after applying the spatial and temporal proximity score. Red lines correspond to the results of fusion of faced cameras 1 with 2 and 5 with 6. Blue line corresponds to the result of fusion of faced cameras 3 and 4.

The corresponding unique IDs are:

- Referee: 200
- Linesmen: 201, 202
- White team goalkeeper: 1
- Blue team goalkeeper: 105
- White team players: 5, 6, 8, 9, 10, 11, 14, 20, 21, 26
- Blue team players: 104, 107, 108, 113, 114, 120, 125, 126, 128, 155

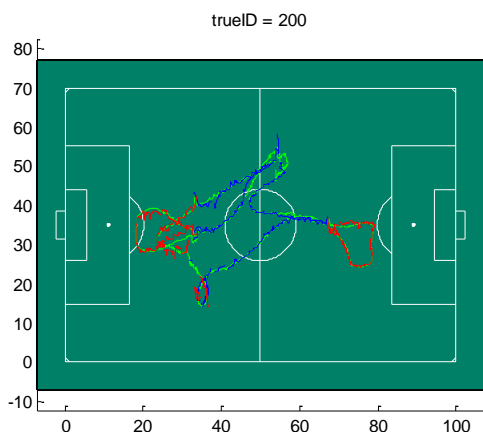


Figure E-1: Trajectory for uniqueID = 200

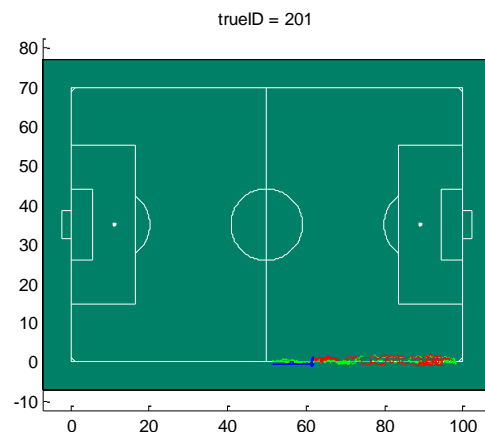


Figure E-2: Trajectory for uniqueID = 201

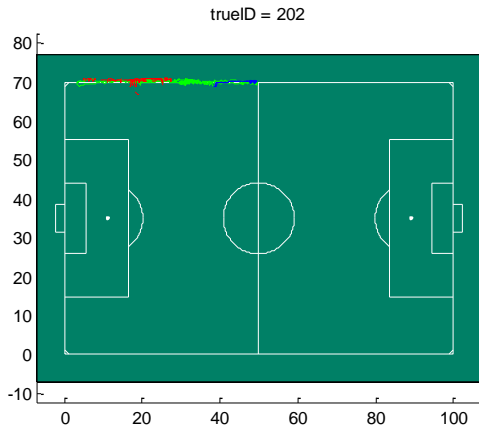


Figure E-3: Trajectory for uniqueID = 202

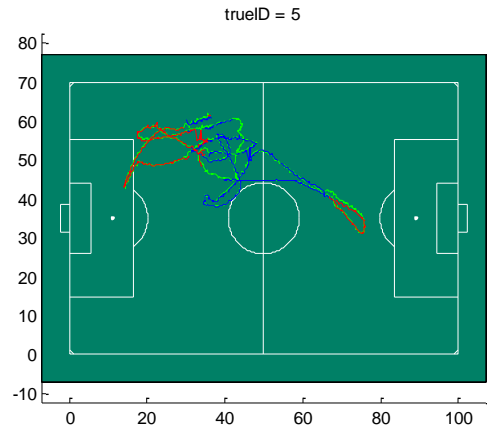


Figure E-6: Trajectory for uniqueID = 5

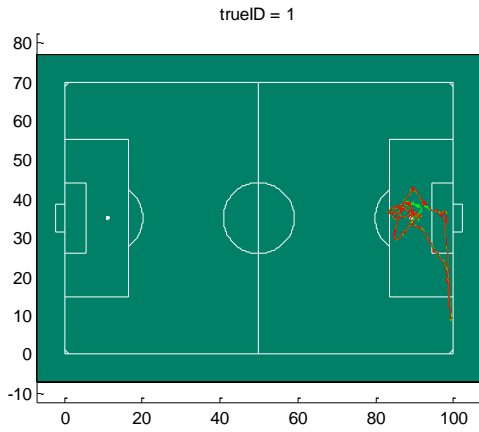


Figure E-4: Trajectory for uniqueID = 1

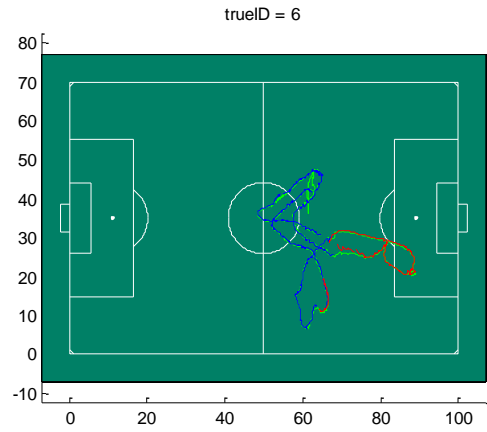


Figure E-7: Trajectory for uniqueID = 6

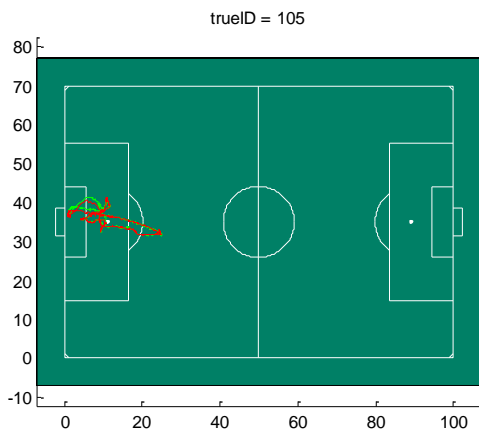


Figure E-5: Trajectory for uniqueID = 105

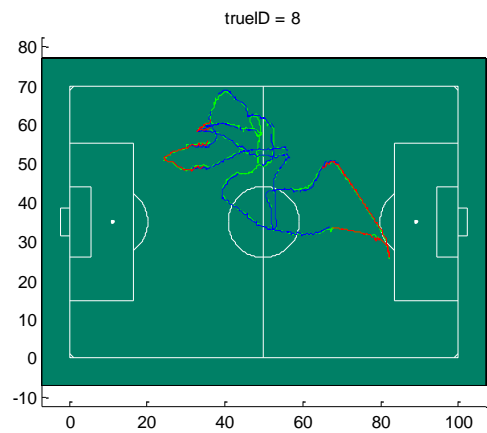


Figure E-8: Trajectory for uniqueID = 8

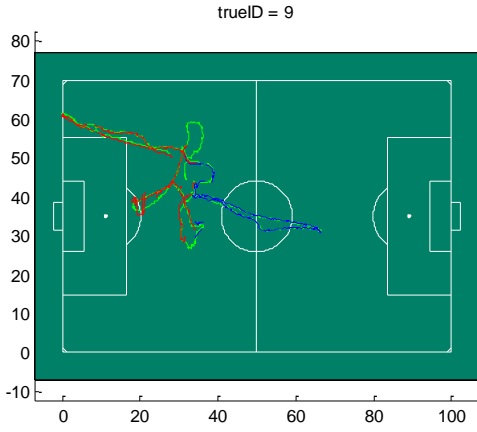


Figure E-9: Trajectory for uniqueID = 9

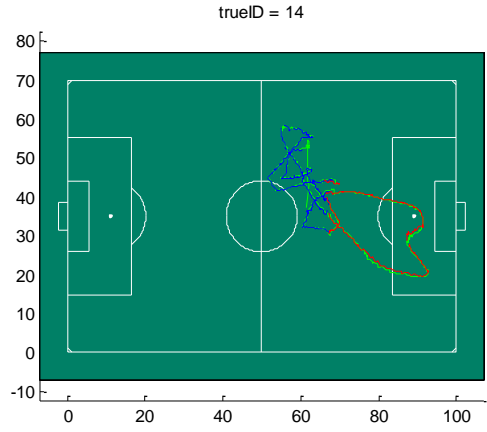


Figure E-12: Trajectory for uniqueID = 14

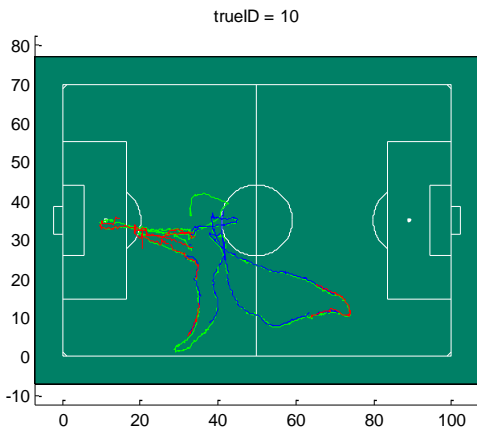


Figure E-10: Trajectory for uniqueID = 10

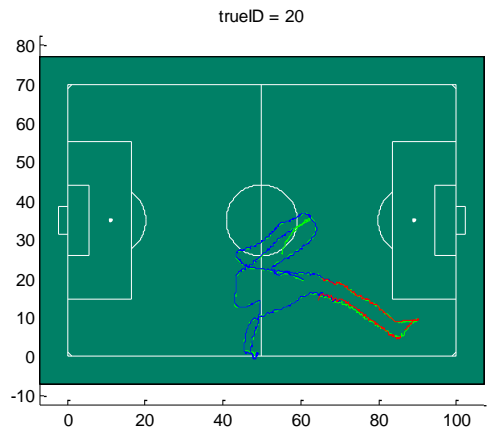


Figure E-13: Trajectory for uniqueID = 20

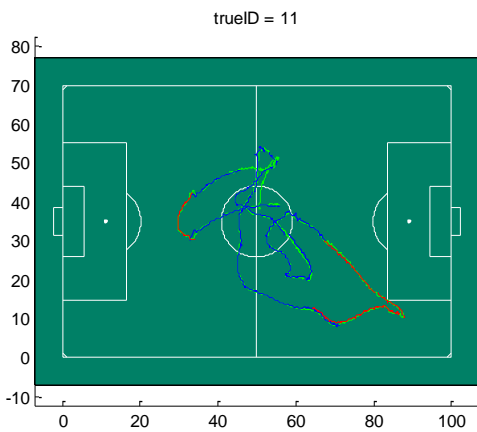


Figure E-11: Trajectory for uniqueID = 11

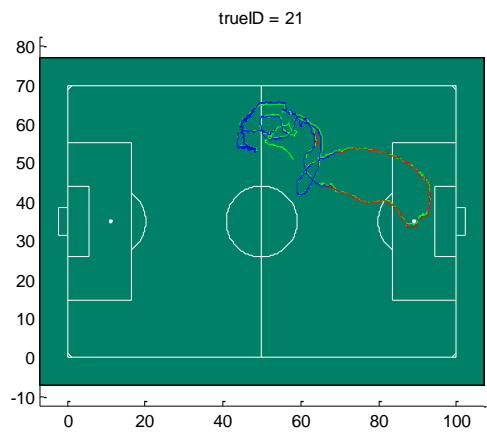


Figure E-14: Trajectory for uniqueID = 21

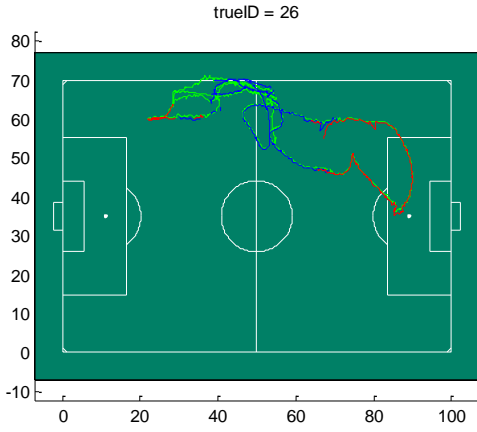


Figure E-15: Trajectory for uniqueID = 26

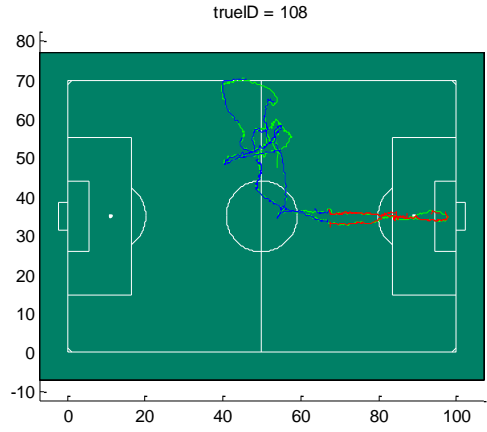


Figure E-18: Trajectory for uniqueID = 108

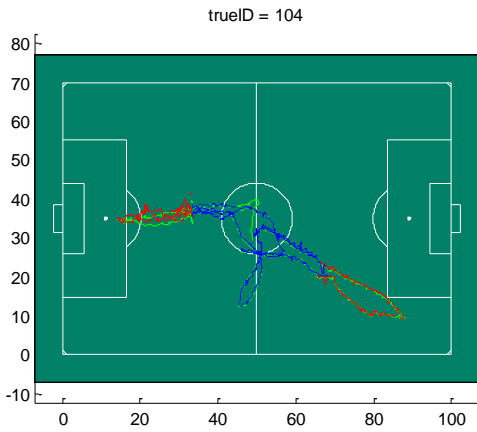


Figure E-16: Trajectory for uniqueID = 104

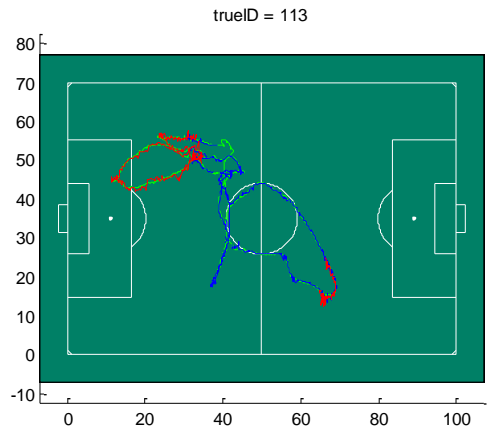


Figure E-19: Trajectory for uniqueID = 113

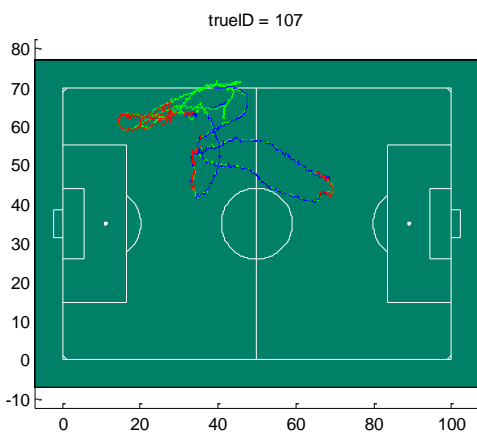


Figure E-17: Trajectory for uniqueID = 107

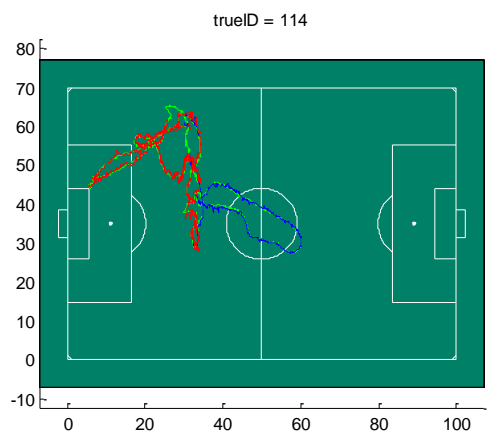


Figure E-20: Trajectory for uniqueID = 114

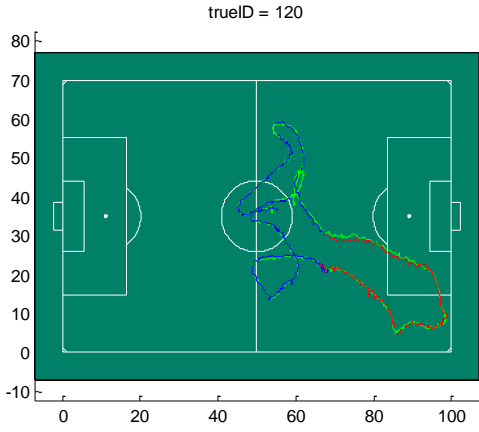


Figure E-21: Trajectory for uniqueID = 120

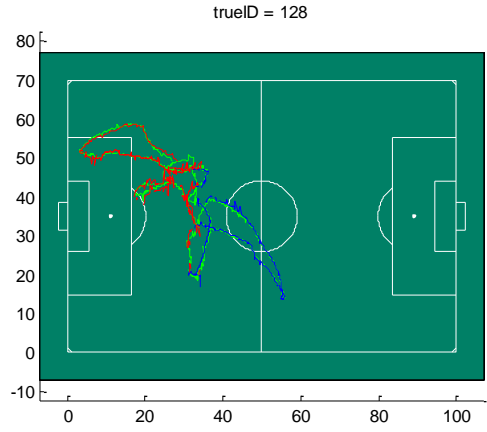


Figure E-24: Trajectory for uniqueID = 128

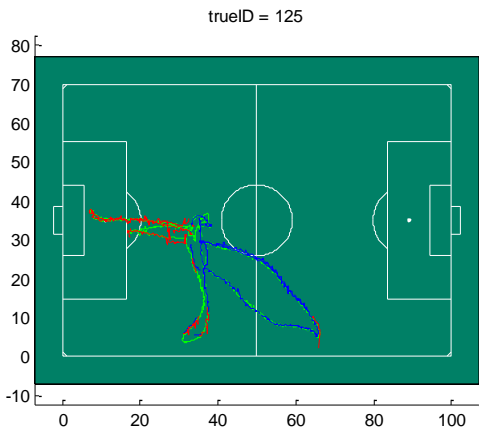


Figure E-22: Trajectory for uniqueID = 125

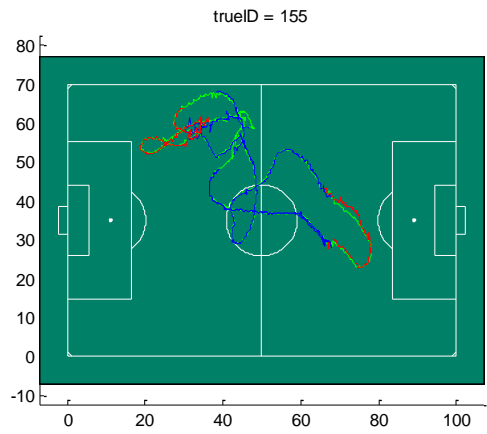


Figure E-25: Trajectory for uniqueID = 155

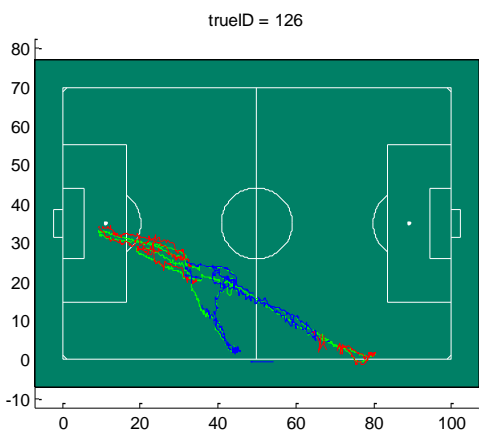


Figure E-23: Trajectory for uniqueID = 126

F.Results of team sports using ground truth tracking

On this annex the results from the fusion in team sports are presented²⁴.

Results from fusion between facing cameras are shown in figures F-1, F-2, F-3 and F-4. Figure F-4 is obtained joining the results obtained from facing cameras fusion. The result is not the average from the 3 pairs because there is not the same number of fusions in each pair. Figure F-5 shows the results of fusing the ideal resulting tracking from facing cameras, it is, fusion of resulting tracking between cameras 3 and camera 4 with resulting tracking between camera 1 with camera 2, and camera 5 with camera 6. The overlapping areas between the different regions are shown in figure 5-7.

Legend:

- GT1 corresponds to the basic fusion.
- GT2 corresponds to the fusion in which is added the improvement of color.
- GT3 corresponds to the fusion in which is added the improvement of color and adjusted the homography of cameras 5 and 6.

²⁴ Results with higher resolution in the URL:

<http://www-vpu.eps.uam.es/publications/DetectionAndTrackingInMulticameraSportsVideo/>

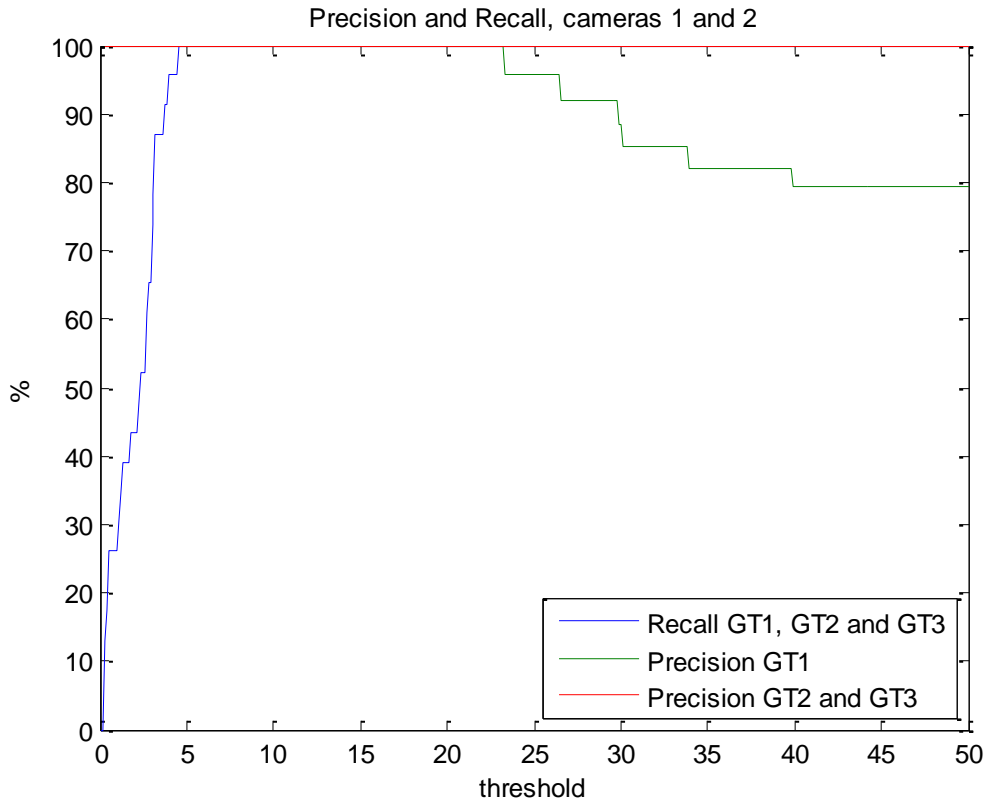


Figure F-1: Precision and Recall from fusion of cameras 1 and 2

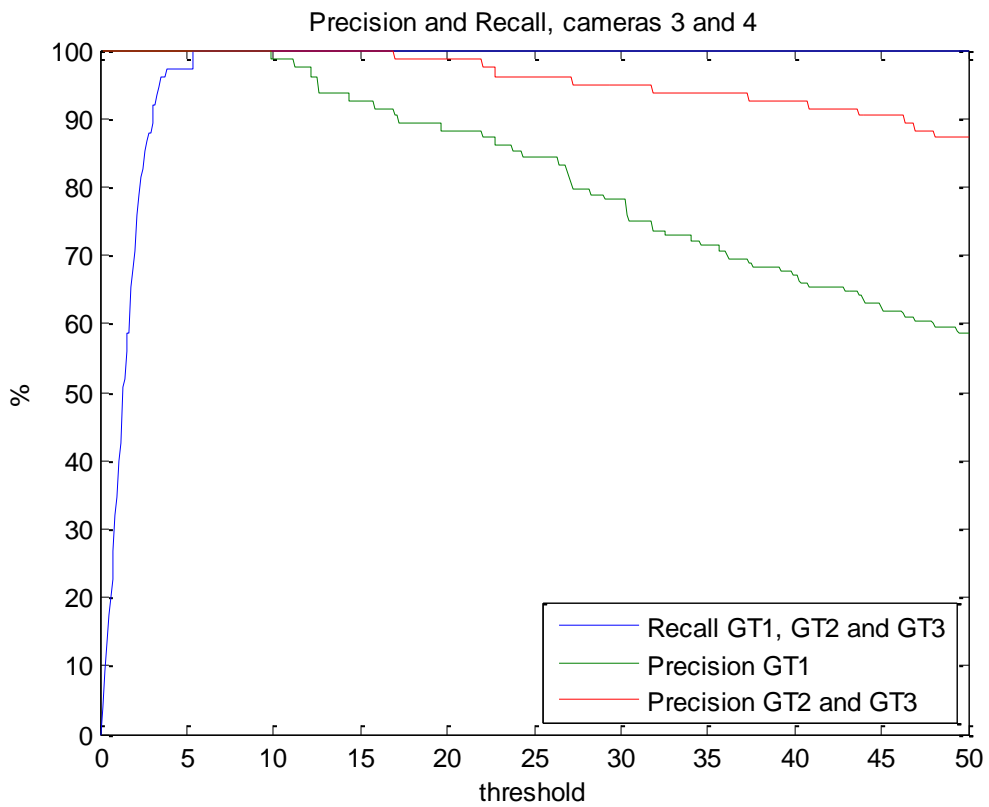


Figure F-2: Precision and Recall from fusion of cameras 3 and 4

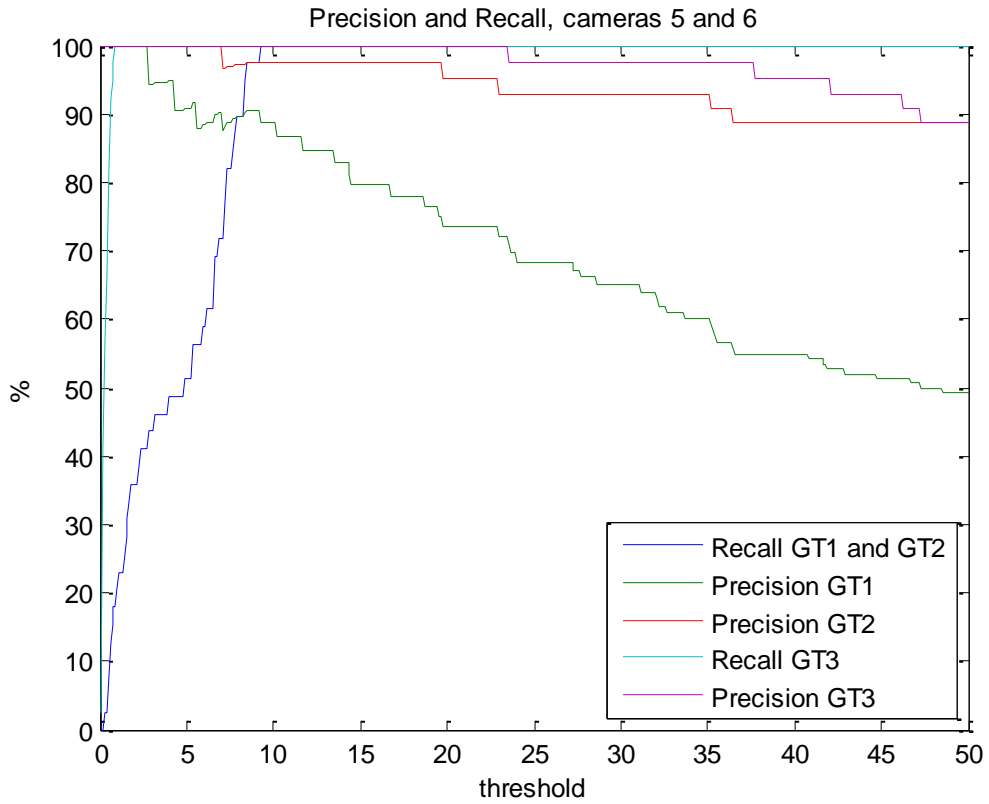


Figure F-3: Precision and Recall from fusion of cameras 5 and 6

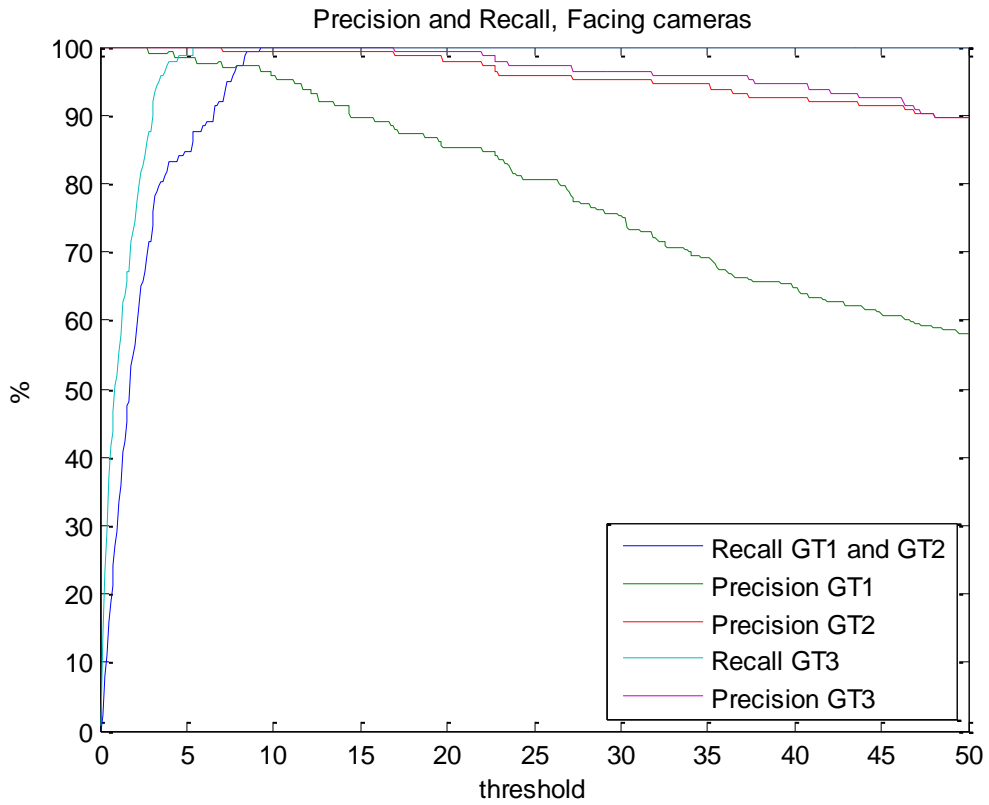


Figure F-4: Precision and Recall from fusion of facing cameras

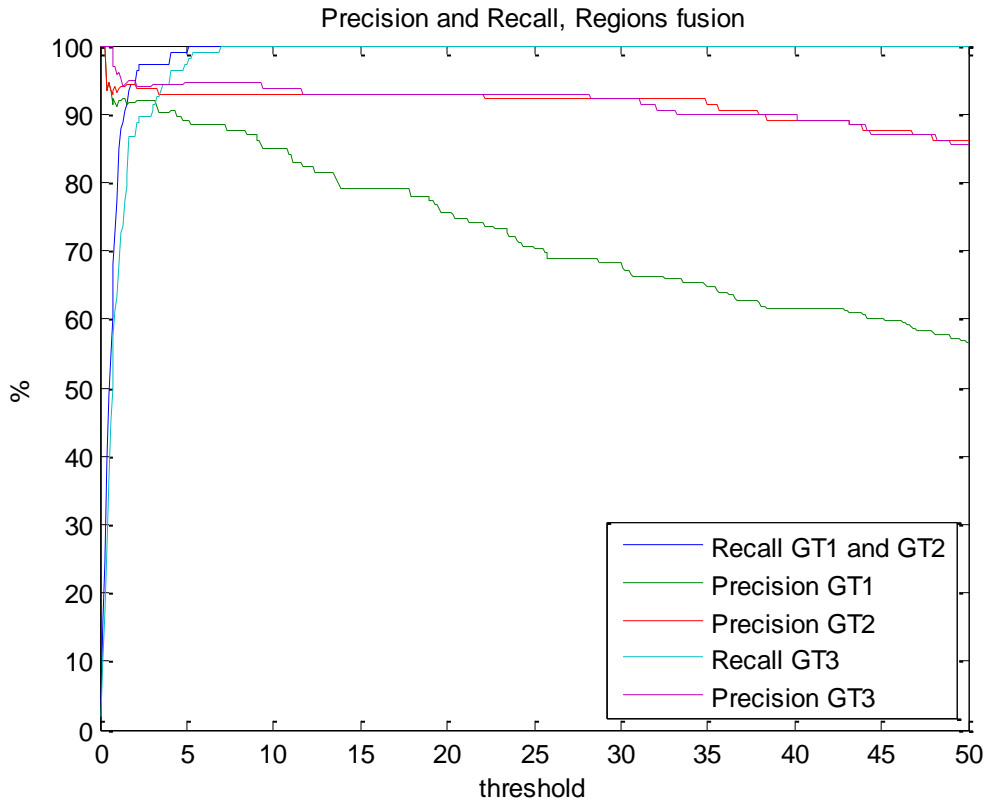


Figure F-5: Precision and Recall from fusion of the different resulting regions

G. Results of team sports using base system tracking

On this annex the results from the fusion in team sports are presented²⁵.

Results from fusion between facing cameras are shown in figures G-1, G-2, G-3 and G-4. Figure G-4 is obtained joining the results obtained from facing cameras fusion. The result is not the average from the 3 pairs because there is not the same number of fusions in each pair. Figure G-5 shows the results of fusing the ideal resulting tracking from facing cameras, it is, fusion of resulting tracking between cameras 3 and camera 4 with resulting tracking between camera 1 with camera 2, and camera 5 with camera 6. The overlapping areas between the different regions are shown in figure 5-7.

Legend:

- BS1 corresponds to the basic fusion.
- BS2 corresponds to the fusion in which is added the improvement of color.
- BS3 corresponds to the fusion in which is added the improvement of color and adjusted the homography of cameras 5 and 6.

²⁵ Results with higher resolution in the URL:

<http://www-vpu.eps.uam.es/publications/DetectionAndTrackingInMulticameraSportsVideo/>

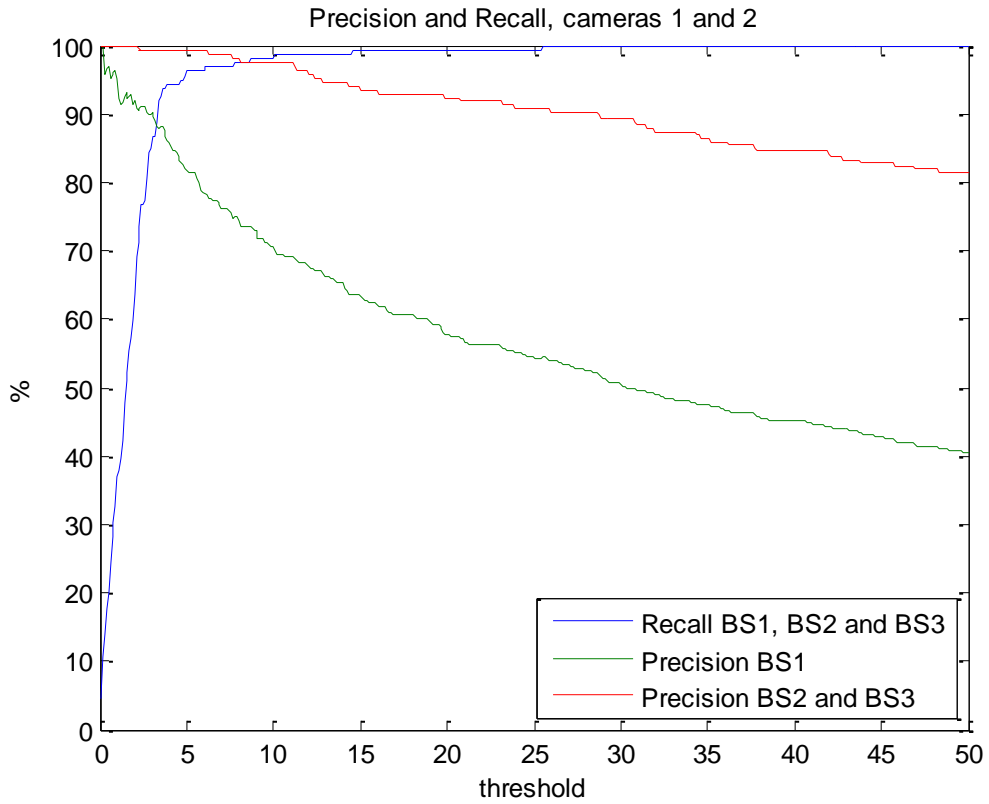


Figure G-1: Precision and Recall from fusion of cameras 1 and 2

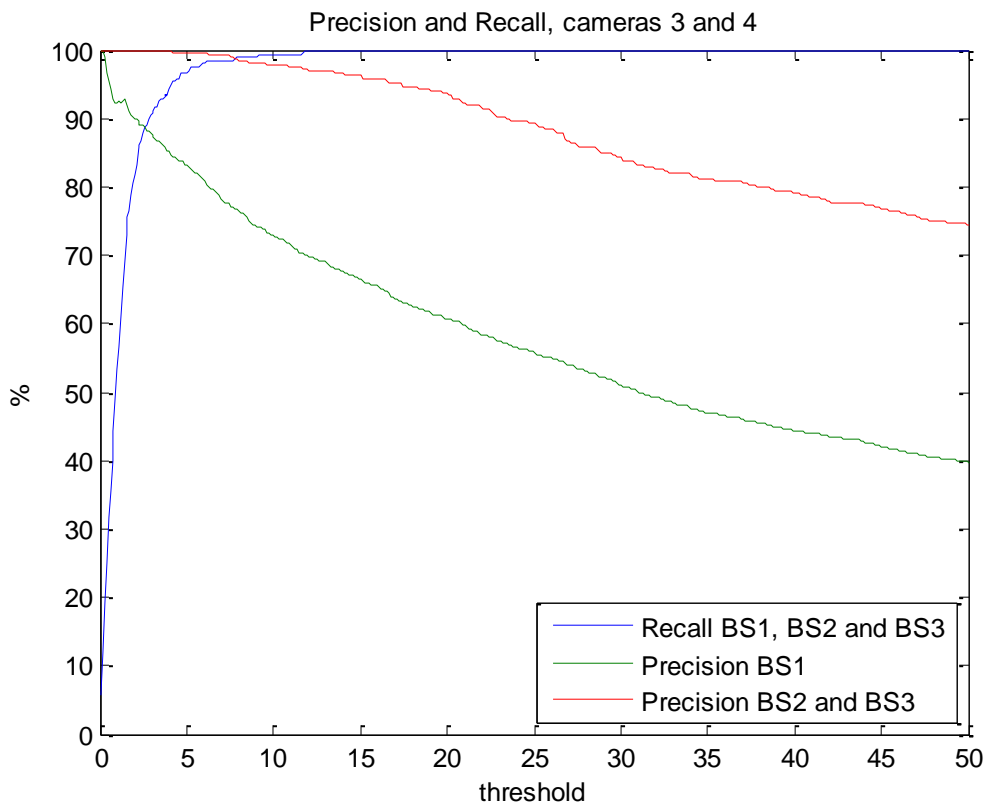


Figure G-2: Precision and Recall from fusion of cameras 3 and 4

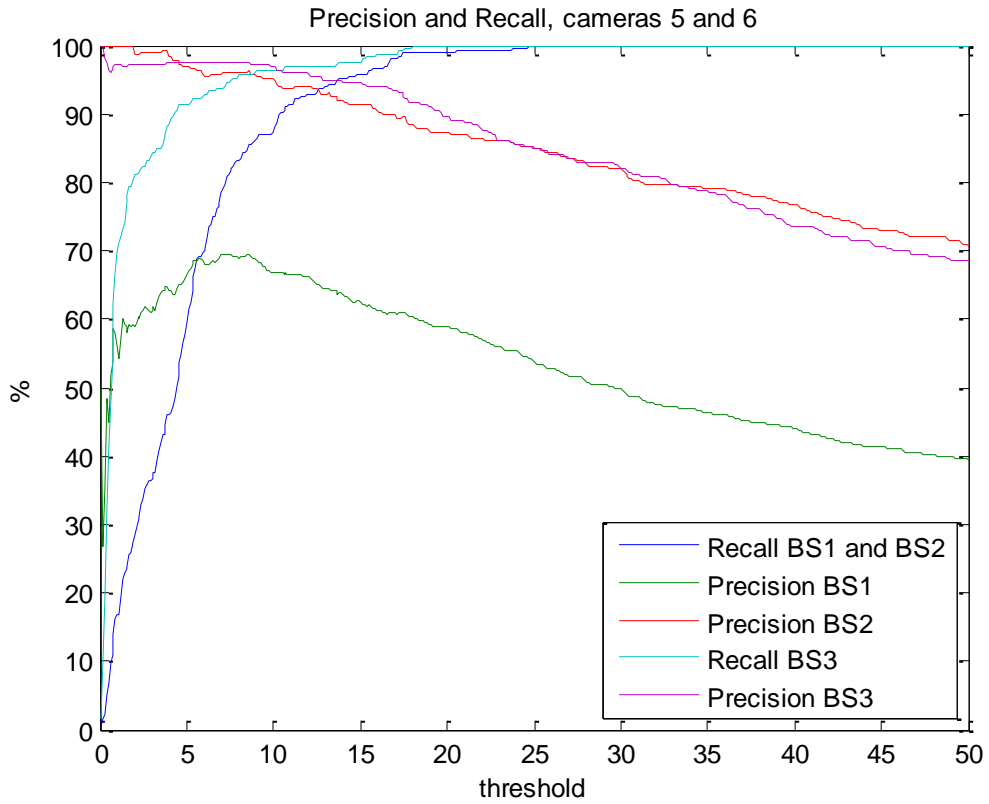


Figure G-3: Precision and Recall from fusion of cameras 5 and 6

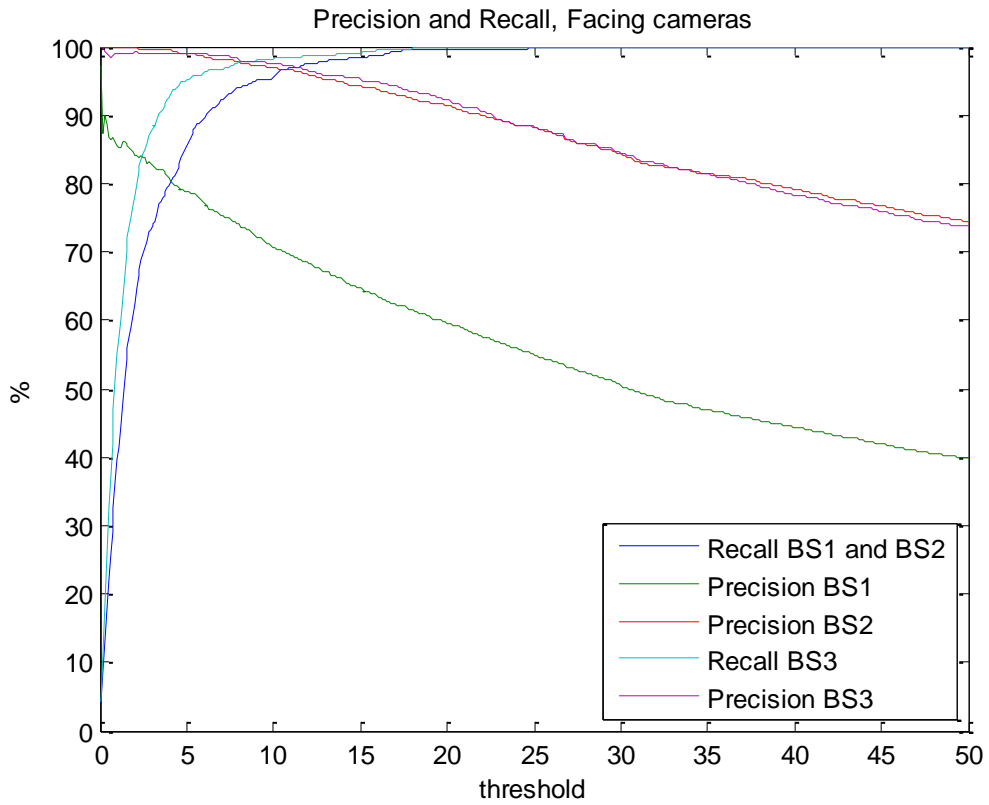


Figure G-4: Precision and Recall from fusion of facing cameras

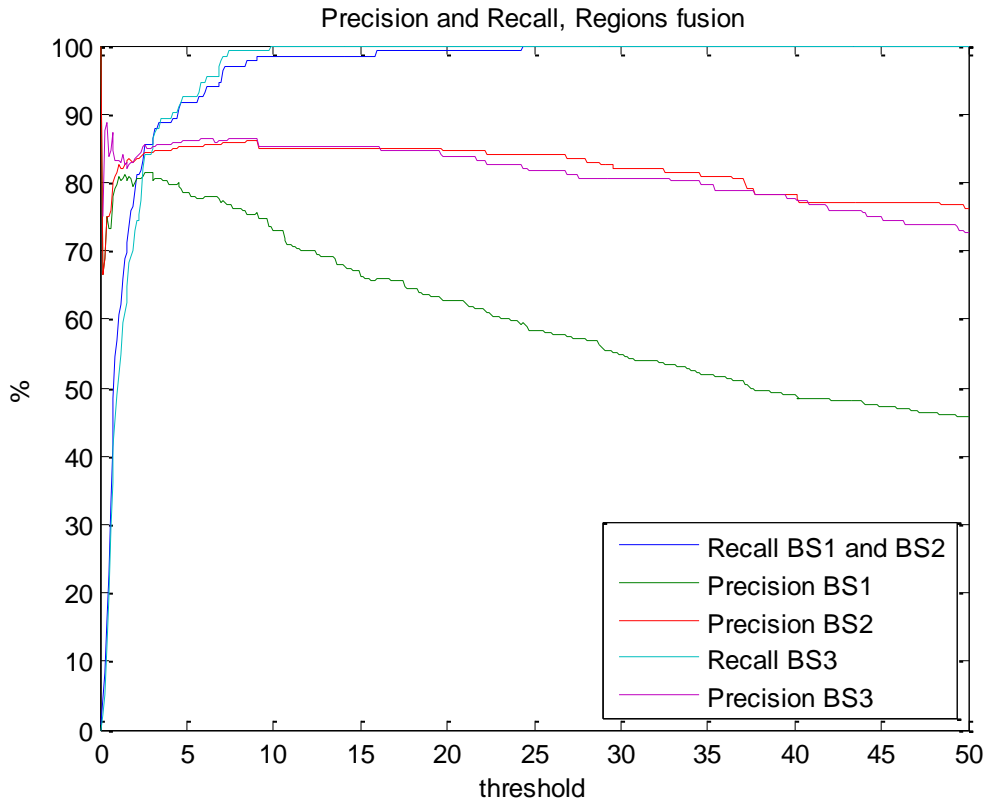


Figure G-5: Precision and Recall from fusion of the different resulting regions

H. Results of team sports. Contrast between ground truth tracking and base system tracking

On this annex the results from the basic fusion in team sports are presented²⁶.

Results from fusion between facing cameras are shown in figures H-1, H-2, H-3 and H-4. Figure H-4 is obtained joining the results obtained from facing cameras fusion. The result is not the average from the 3 pairs because there is not the same number of fusions in each pair. Figure H-5 shows the results of fusion the ideal resulting tracking from facing cameras, it is, fusion of resulting tracking between cameras 3 and camera 4 with resulting tracking between camera 1 with camera 2, and camera 5 with camera 6. The overlapping areas between the different regions are shown in figure 5-4.

Meaning of legend (GT1, GT2, GT3, BS1, BS2 and BS3) is depicted in Annex F and Annex A.

²⁶ Results with higher resolution in the URL:

<http://www-vpu.eps.uam.es/publications/DetectionAndTrackingInMulticameraSportsVideo/>

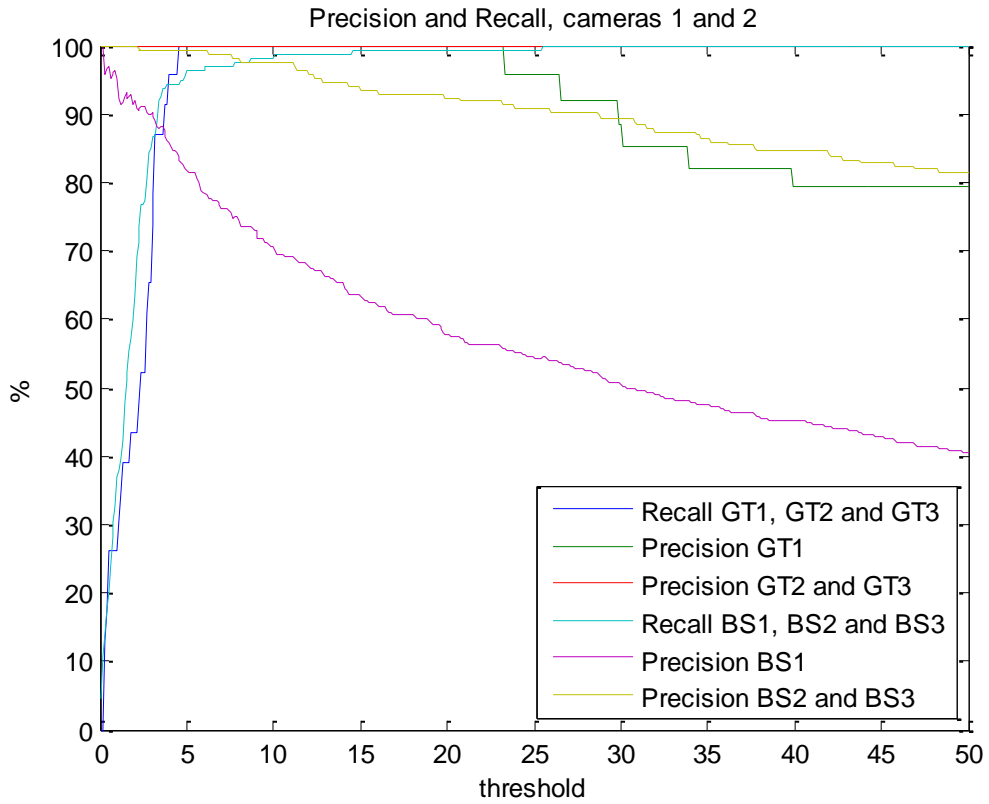


Figure H-1: Precision and Recall from fusion of cameras 1 and 2

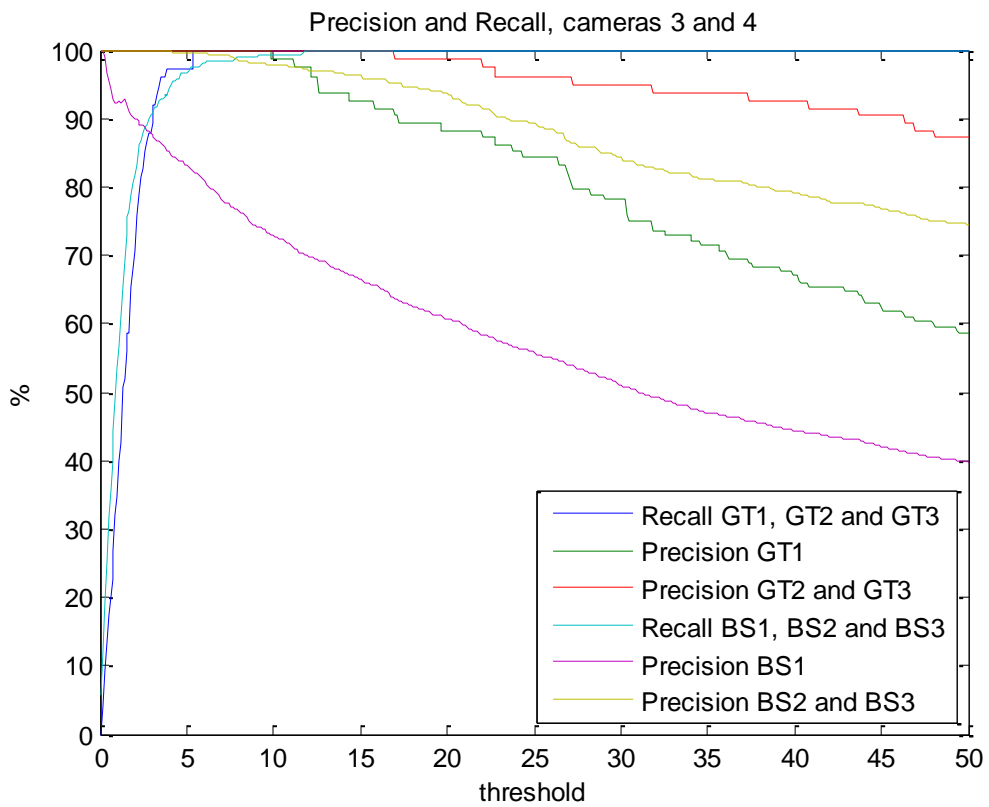


Figure H-2: Precision and Recall from fusion of cameras 3 and 4

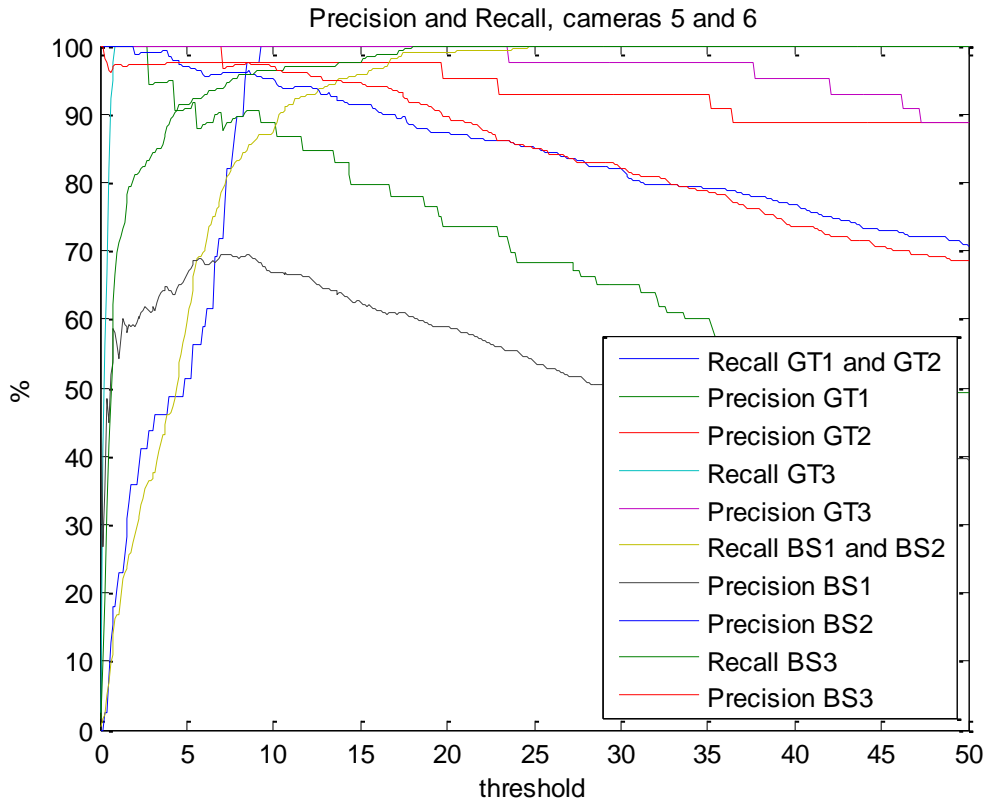


Figure H-3: Precision and Recall from fusion of cameras 5 and 6

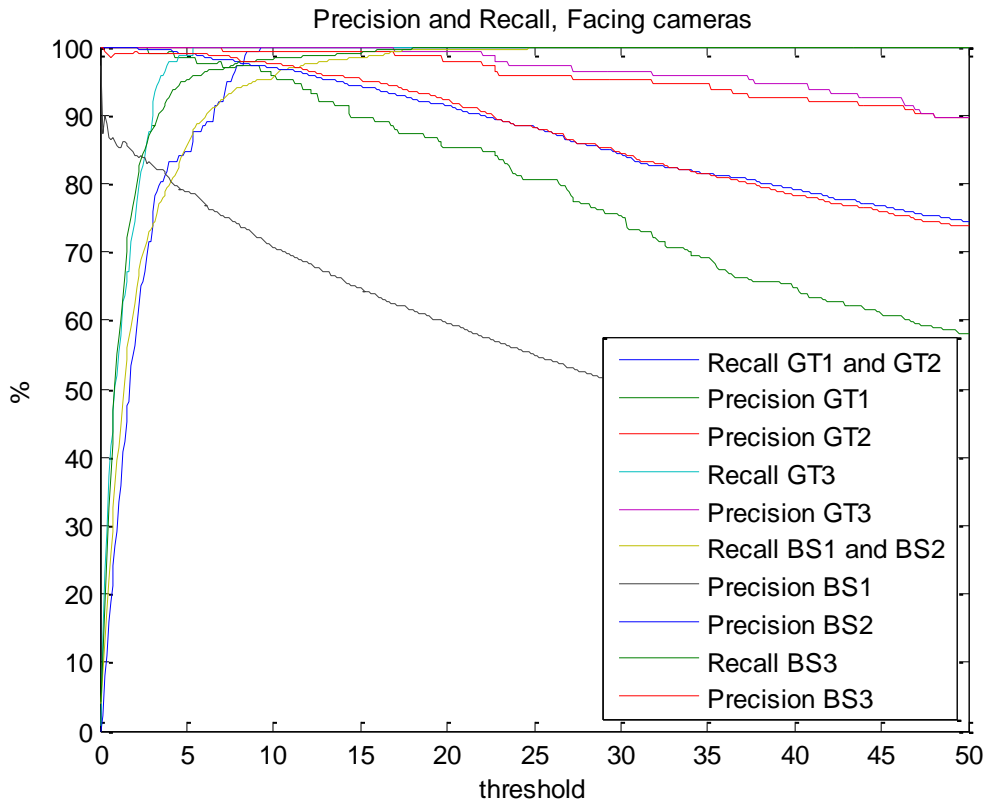


Figure H-4: Precision and Recall from fusion of facing cameras

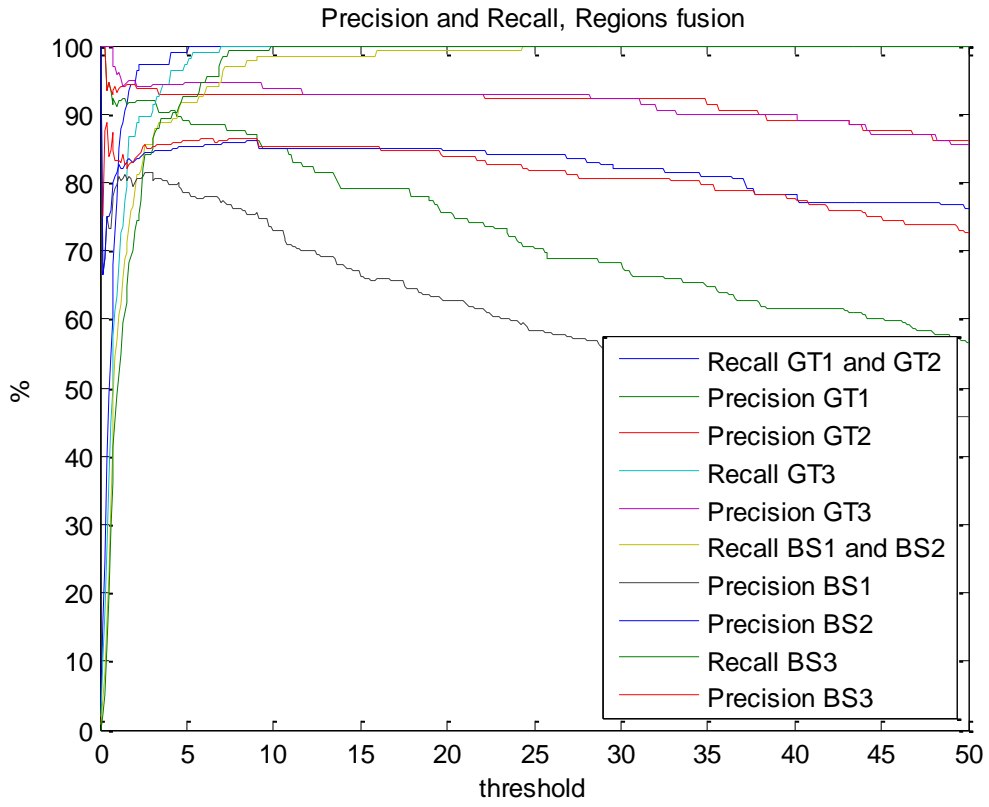


Figure H-5: Precision and Recall from fusion of the different resulting regions

I. Player statistics extracted from the results of the systems

On this annex, the statistics extracted from the developed systems are shown. The presented statistics corresponds to the tennis player and to the 25 tracked people in the football system. The given statistics correspond to the full video.

I.1 Tennis system

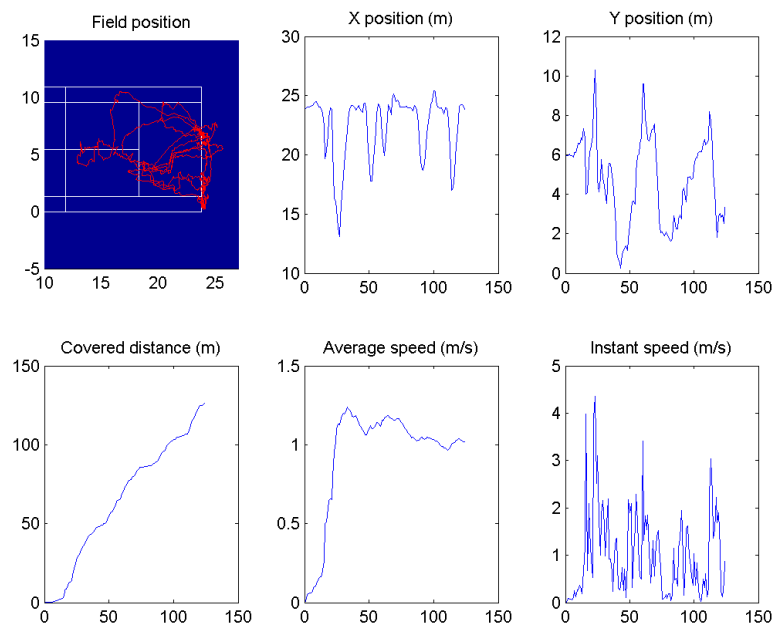


Figure I-1: Statistics for the tennis player

I.2 Football system

The corresponding unique IDs are:

- Referee: 200
- Linesmen: 201, 202
- White team goalkeeper: 1
- Blue team goalkeeper: 105
- White team players: 5, 6, 8, 9, 10, 11, 14, 20, 21, 26
- Blue team players: 104, 107, 108, 113, 114, 120, 125, 126, 128, 155

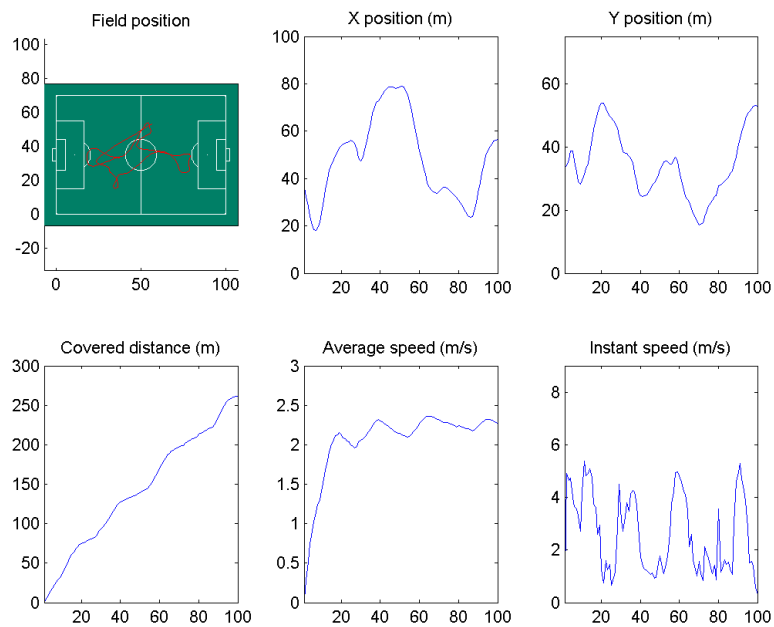


Figure I-2: Statistics for uniqueID = 200

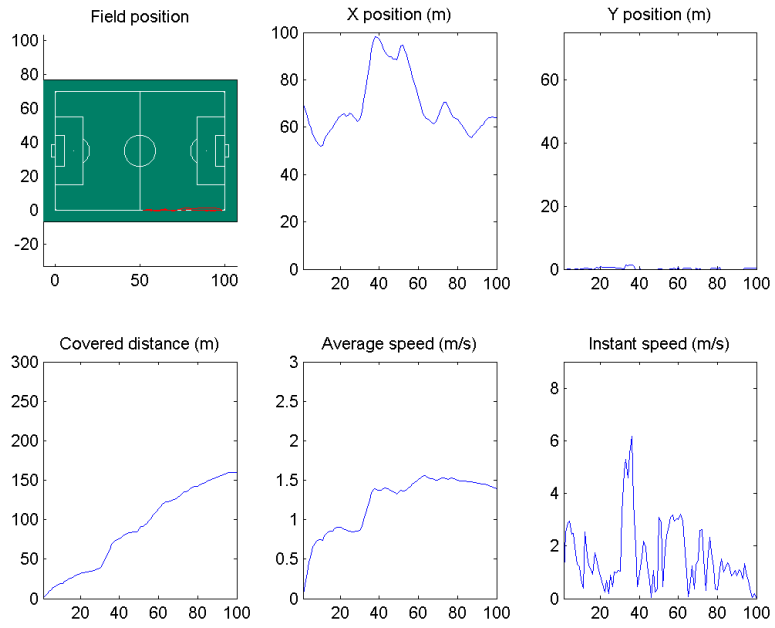


Figure I-3: Statistics for uniqueID = 201

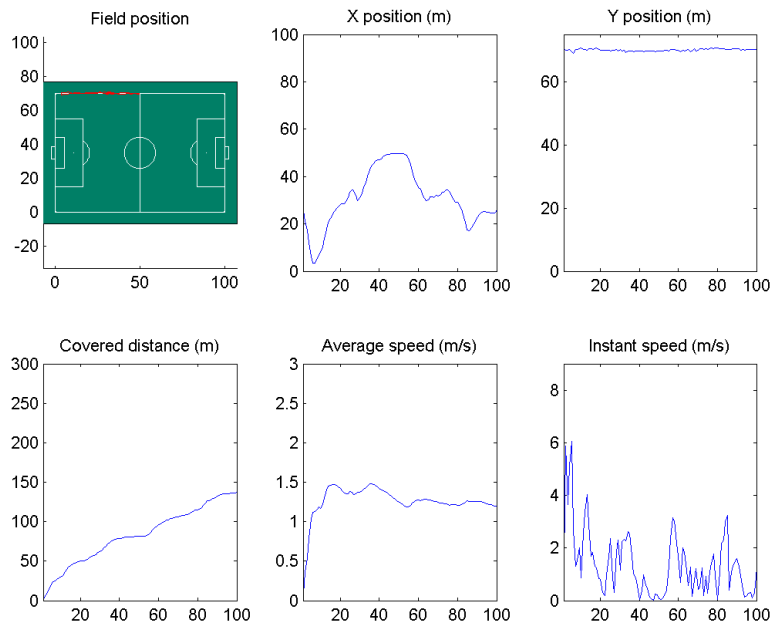


Figure I-4: Statistics for uniqueID = 202

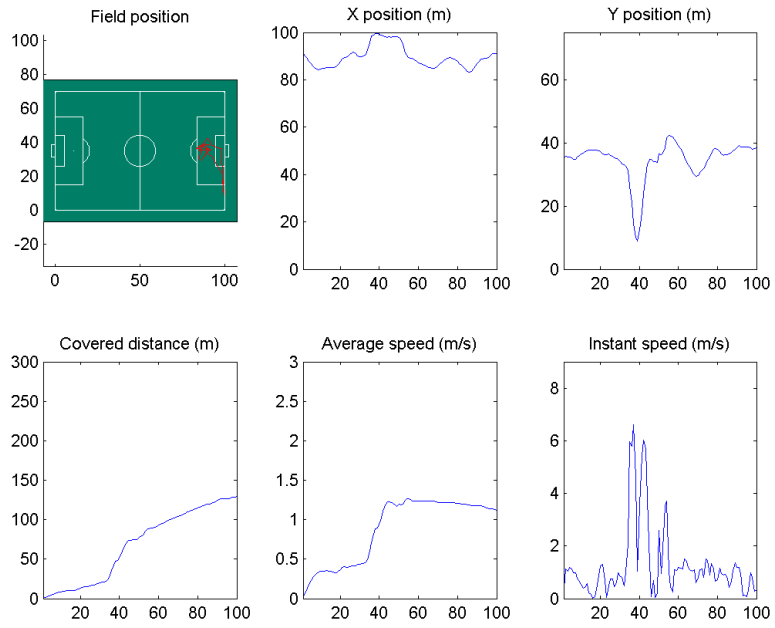


Figure I-5: Statistics for uniqueID = 1

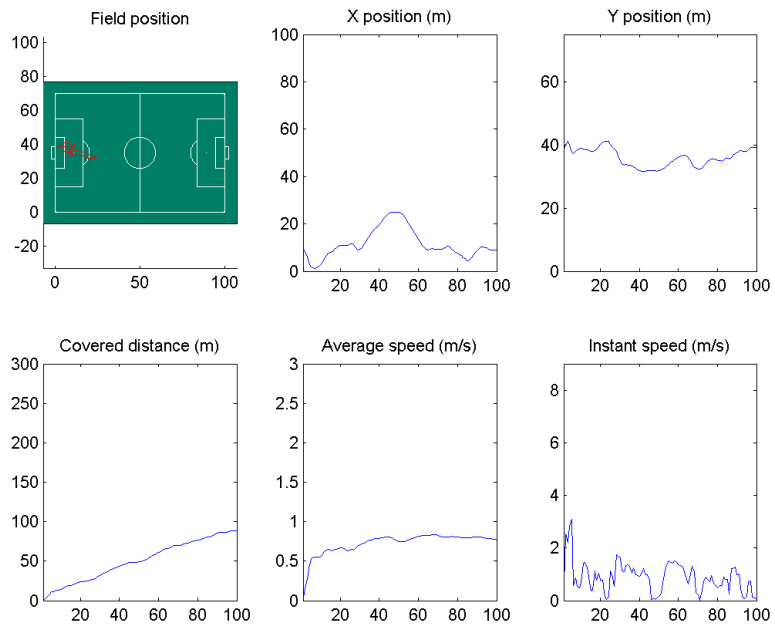


Figure I-6: Statistics for uniqueID = 105

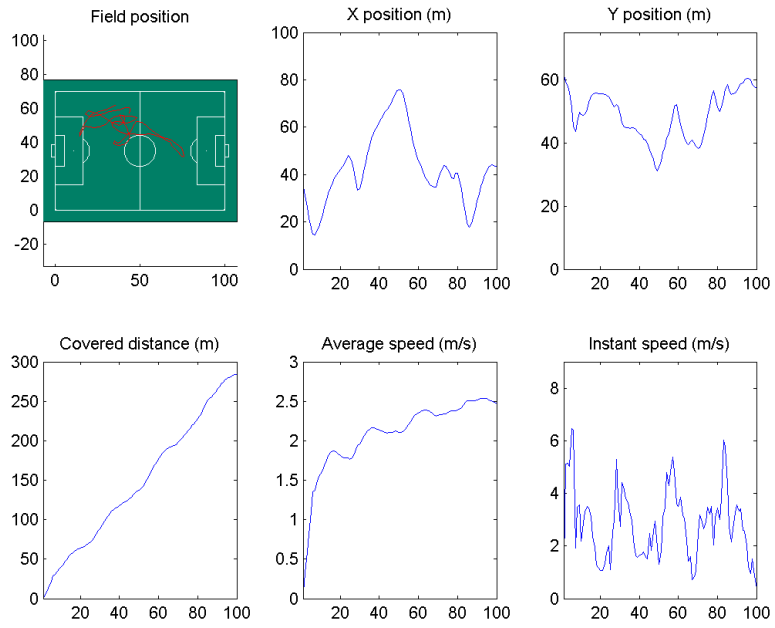


Figure I-7: Statistics for uniqueID = 5

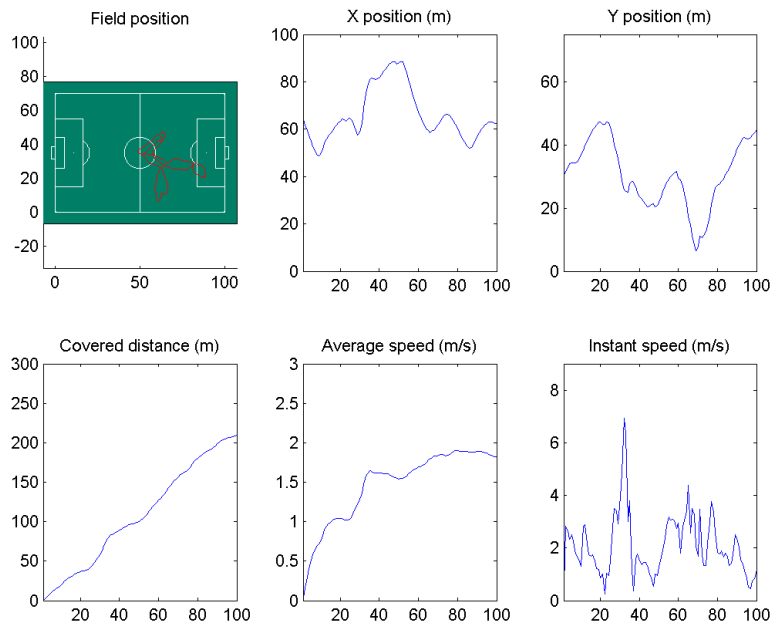


Figure I-8: Statistics for uniqueID = 6

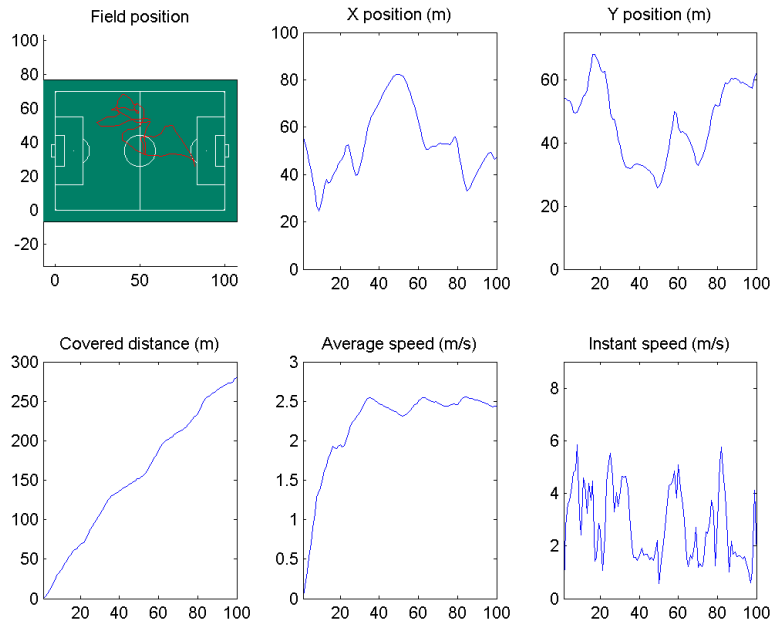


Figure I-9: Statistics for uniqueID = 8

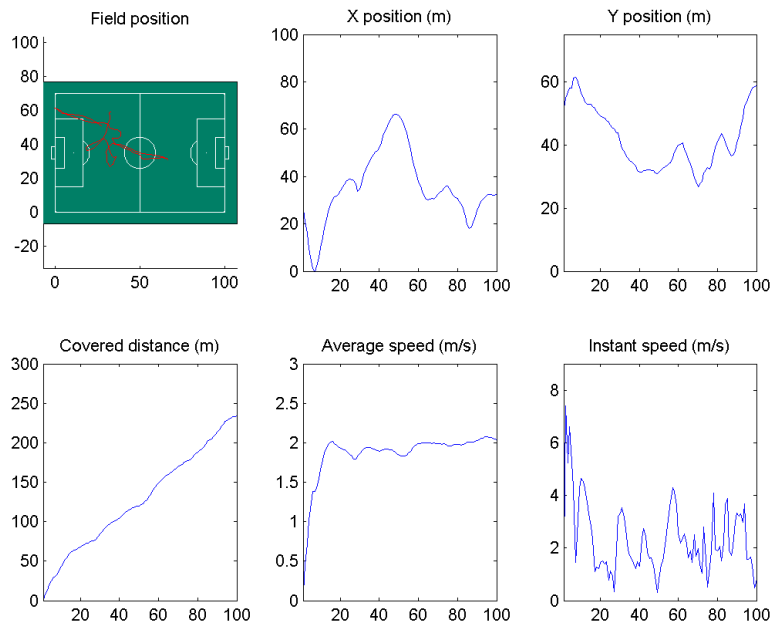


Figure I-10: Statistics for uniqueID = 9

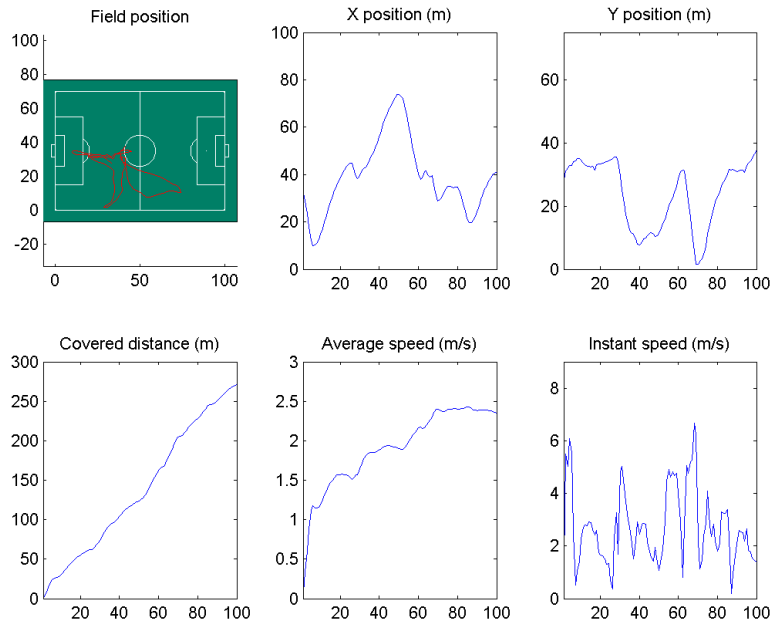


Figure I-11: Statistics for uniqueID = 10

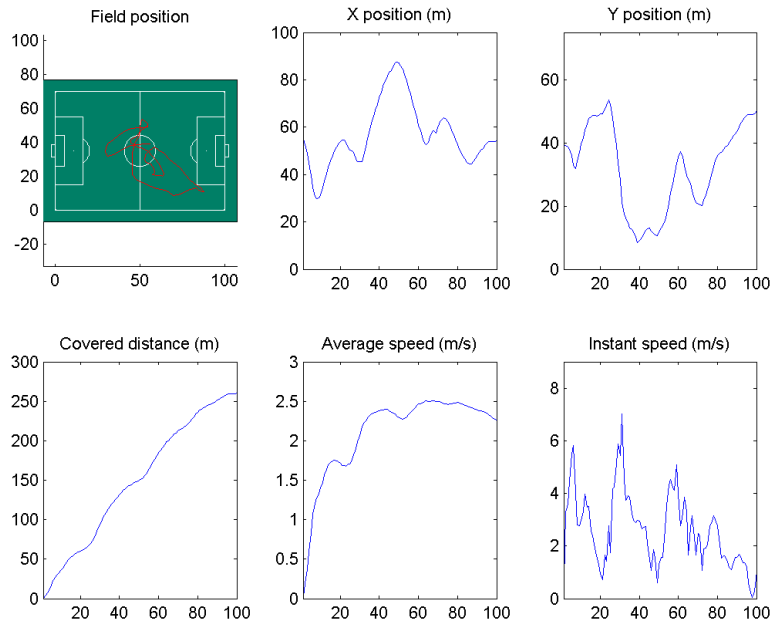


Figure I -12: Statistics for uniqueID = 11

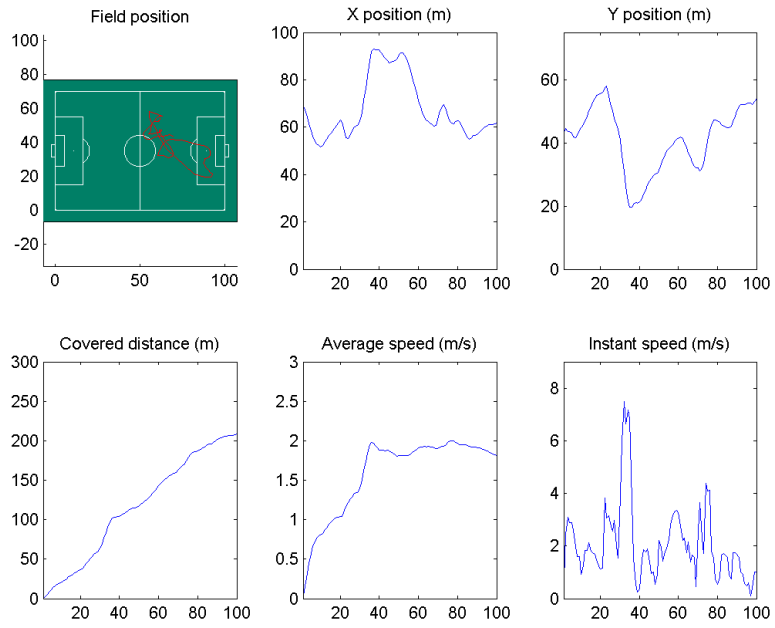


Figure I -13: Statistics for uniqueID = 14

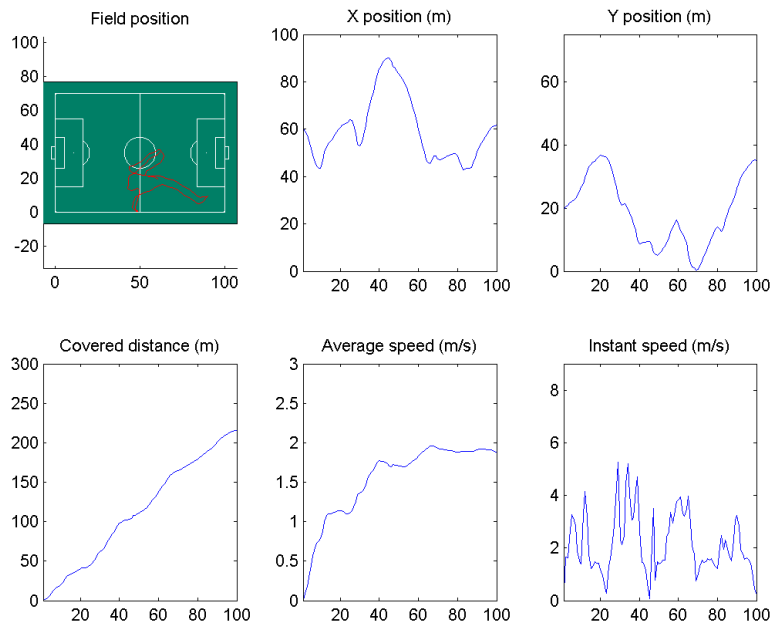


Figure I-14: Statistics for uniqueID = 20

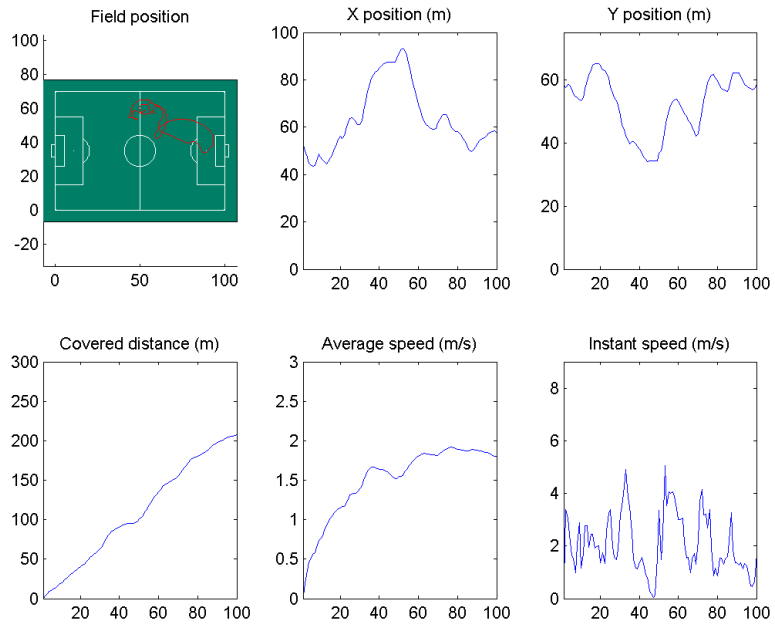


Figure I-15: Statistics for uniqueID = 21

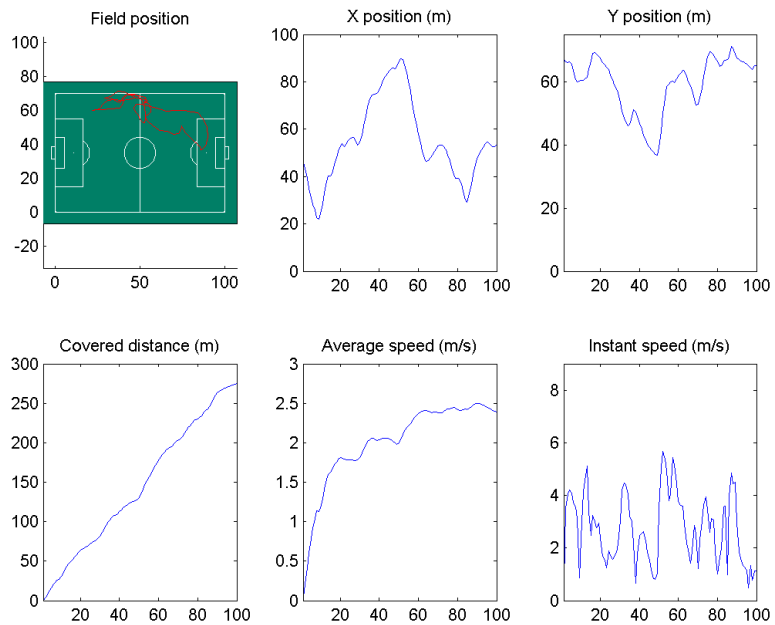


Figure I-16: Statistics for uniqueID = 26

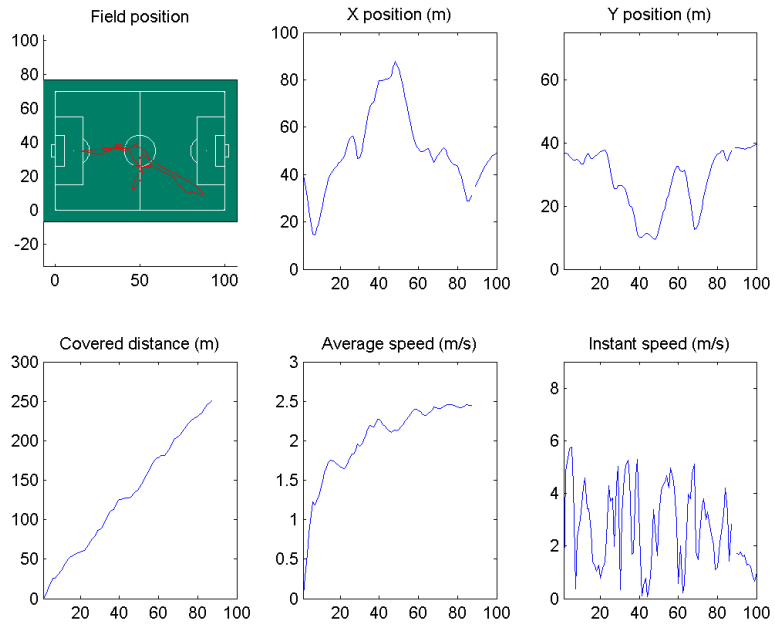


Figure I-17: Statistics for uniqueID = 104

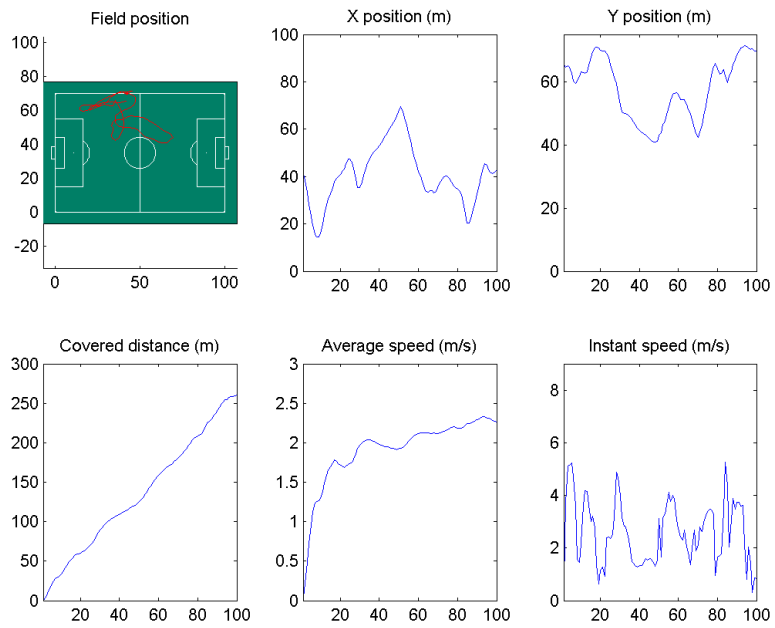


Figure I-18: Statistics for uniqueID = 107

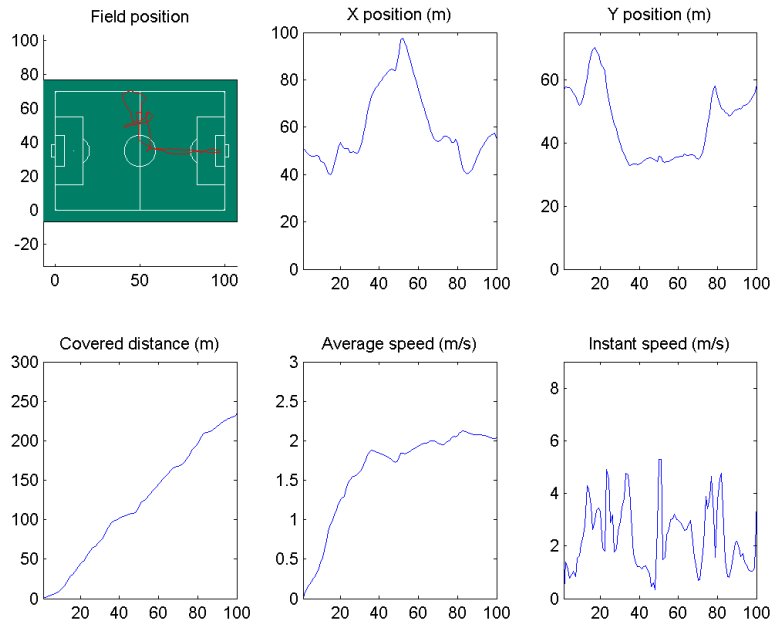


Figure I-19: Statistics for uniqueID = 108

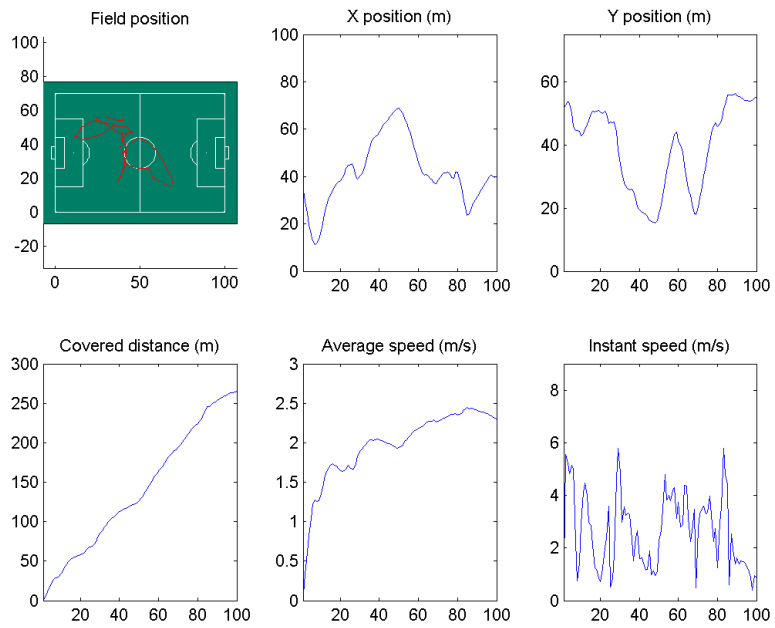


Figure I-20: Statistics for uniqueID = 113

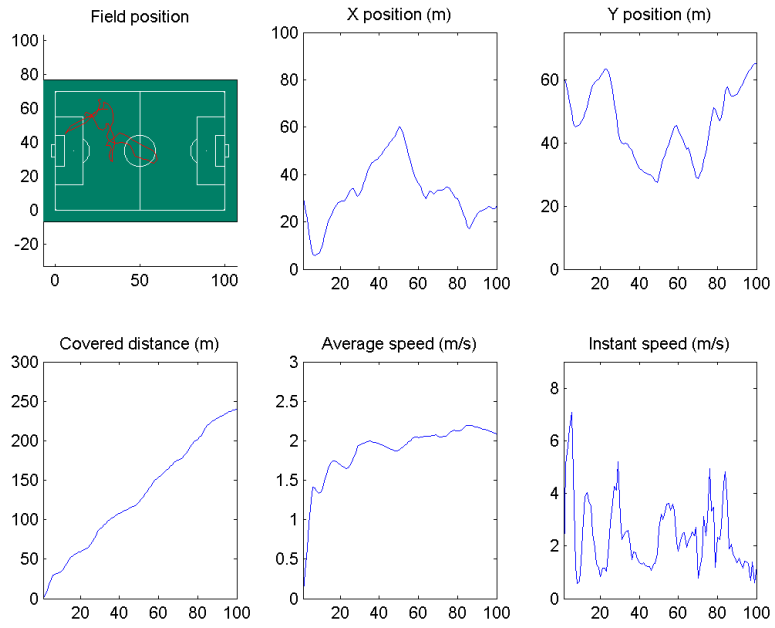


Figure I-21: Statistics for uniqueID = 114

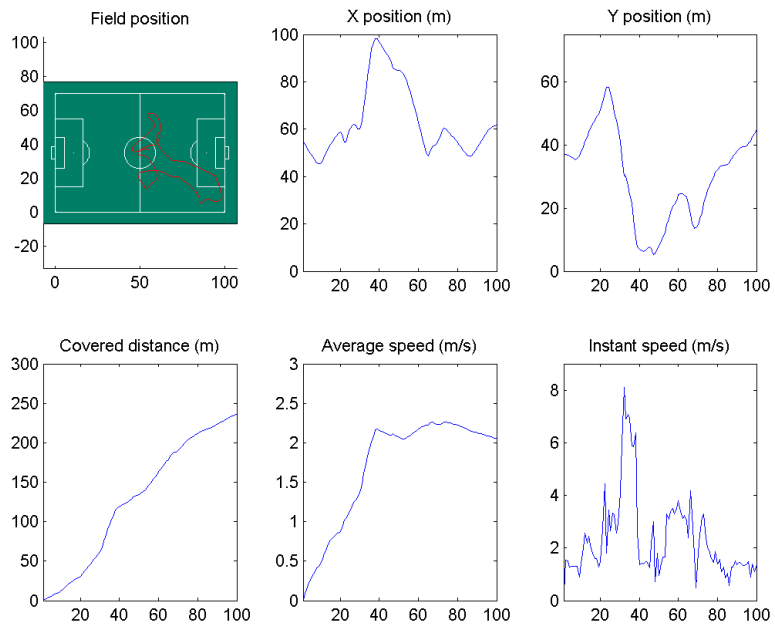


Figure I-22: Statistics for uniqueID = 120

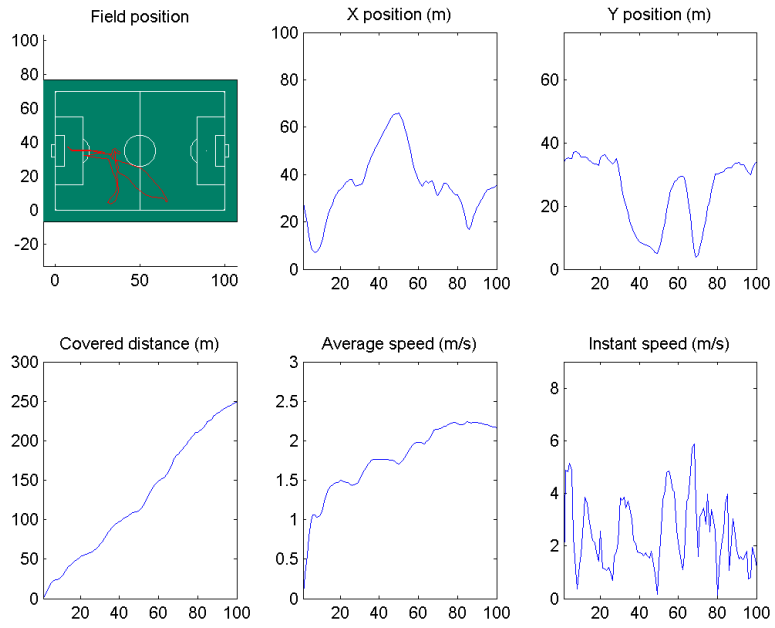


Figure I-23: Statistics for uniqueID = 125

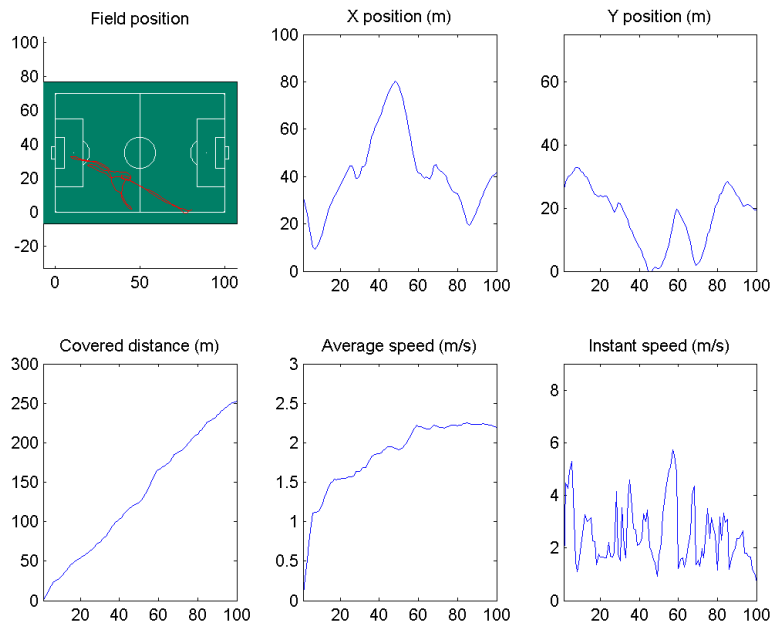


Figure I-24: Statistics for uniqueID = 126

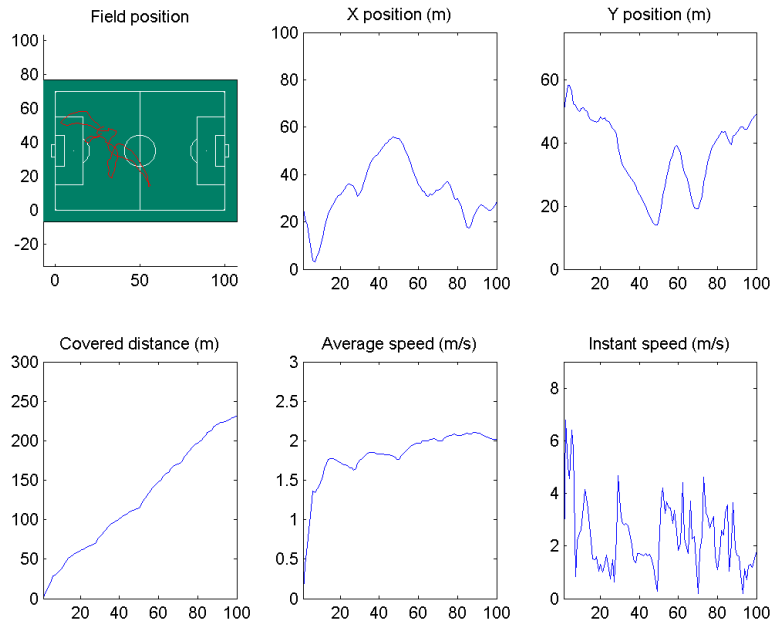


Figure I-25: Statistics for uniqueID = 128

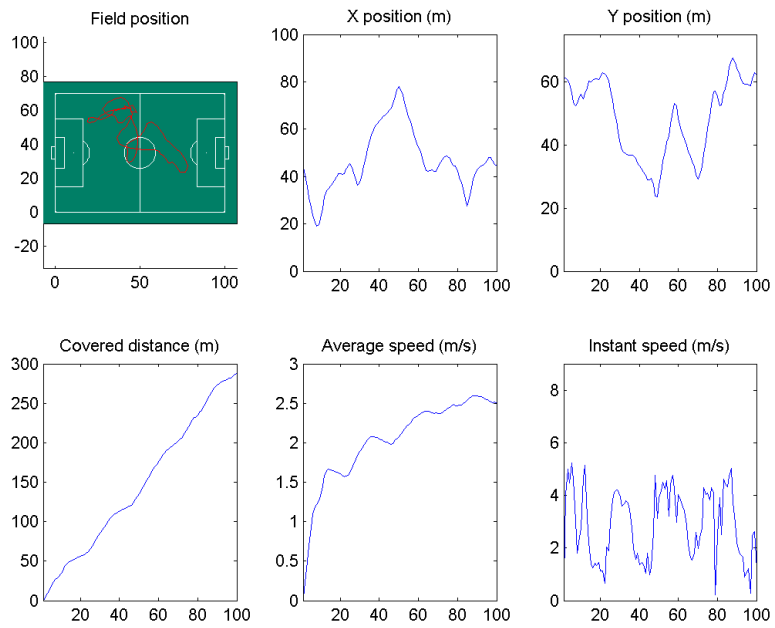


Figure I-26: Statistics for uniqueID = 155

J. Resulting trajectories for the football system

On this annex, four figures of the resulting trajectories for the football system are presented to subjectively evaluate the results and errors. For each scenario (ground truth tracking and base system tracking) the ideal fusion and the experimental fusion are presented. The third method for fusion is used with a threshold value of 5 (heuristically set). In the ideal fusions there are 25 resulting blobs. In the experimental fusions, there are 51 blobs for the ground truth tracking scenario and 402 resulting blobs, for the base system tracking scenario. In the ideal fusion for the base system tracking scenario, all the blobs of each player are displayed in a single color, although all the fusions can not be performed properly due to the absence of tracking in certain cases.

The figures are presented on the following page to see the ideal fusion and the experimental fusion simultaneously.

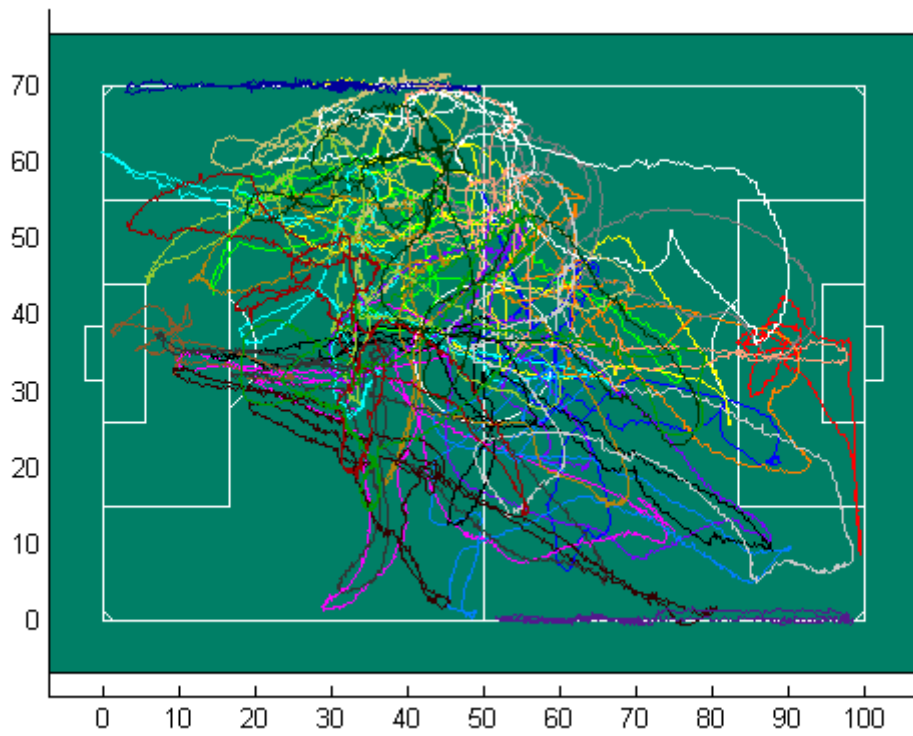


Figure J-1: Resulting trajectories for the ideal fusion in the ground truth tracking scenario

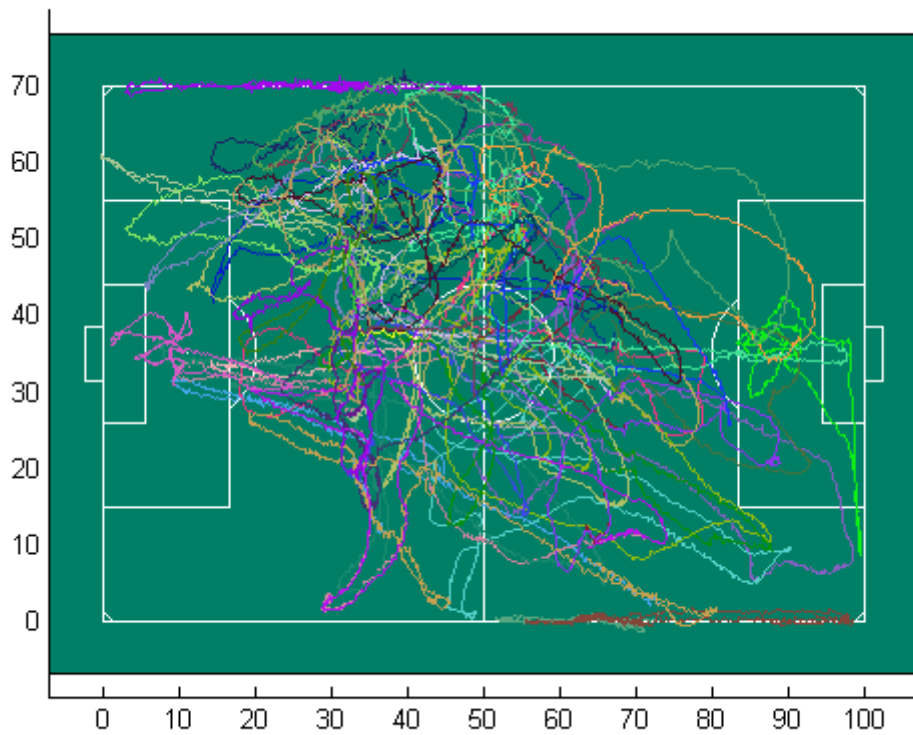


Figure J-2: Resulting trajectories for the experimental fusion in the ground truth tracking scenario

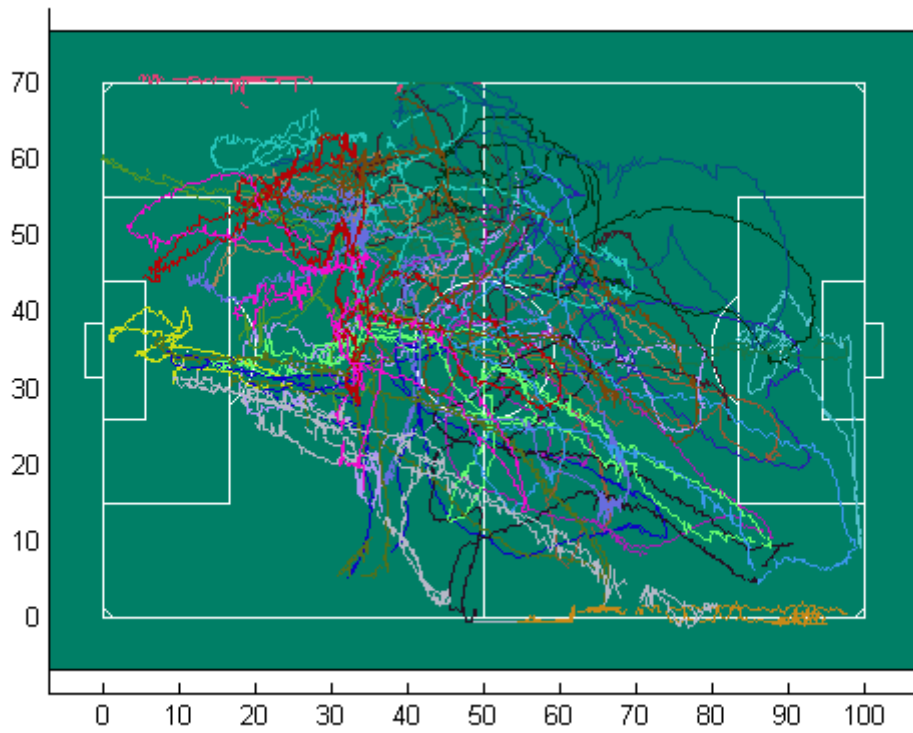


Figure J-3: Resulting trajectories for the ideal fusion in the base system tracking scenario

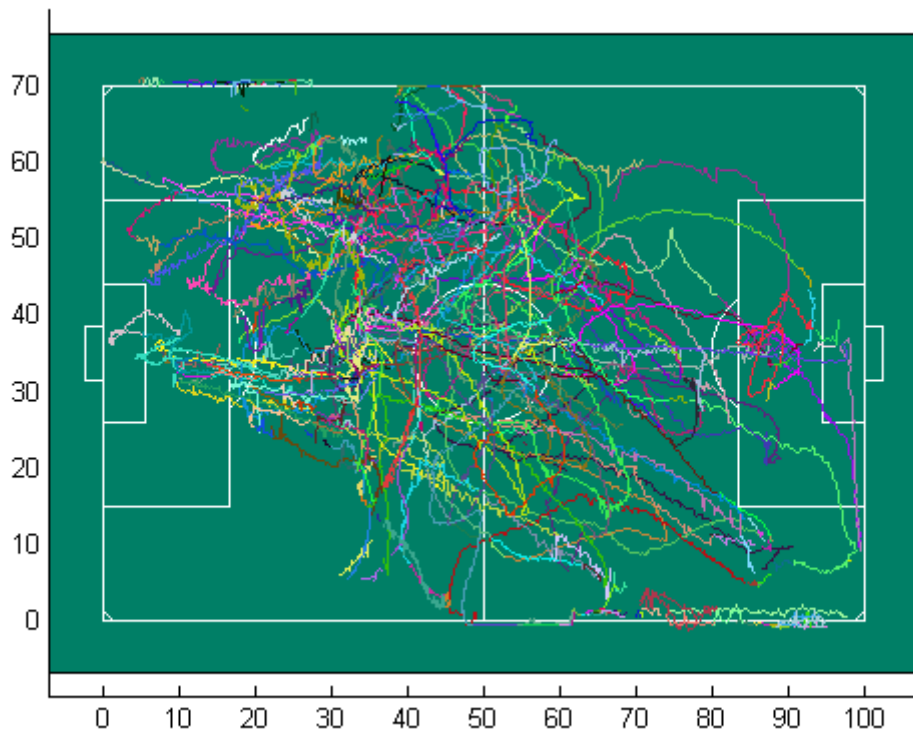


Figure J-4: Resulting trajectories for the experimental fusion in the base system tracking scenario

K. Supervised association of the trajectory fragments of each player for the football system

On this annex, the supervised associated fragments of the trajectories of each player are shown. Next to each image, the complete trajectory of each player, extracted from the ground truth tracking, is shown to compare both results.

The corresponding unique IDs are:

- Referee: 200
- Linesmen: 201, 202
- White team goalkeeper: 1
- Blue team goalkeeper: 105
- White team players: 5, 6, 8, 9, 10, 11, 14, 20, 21, 26
- Blue team players: 104, 107, 108, 113, 114, 120, 125, 126, 128, 155

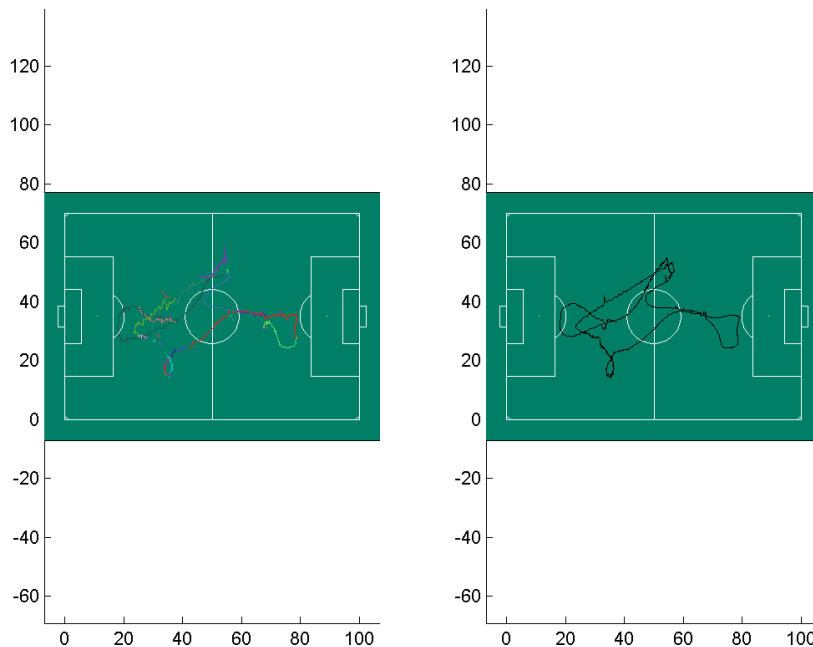


Figure K-1: Supervised associated fragments and complete trajectory for uniqueID = 200

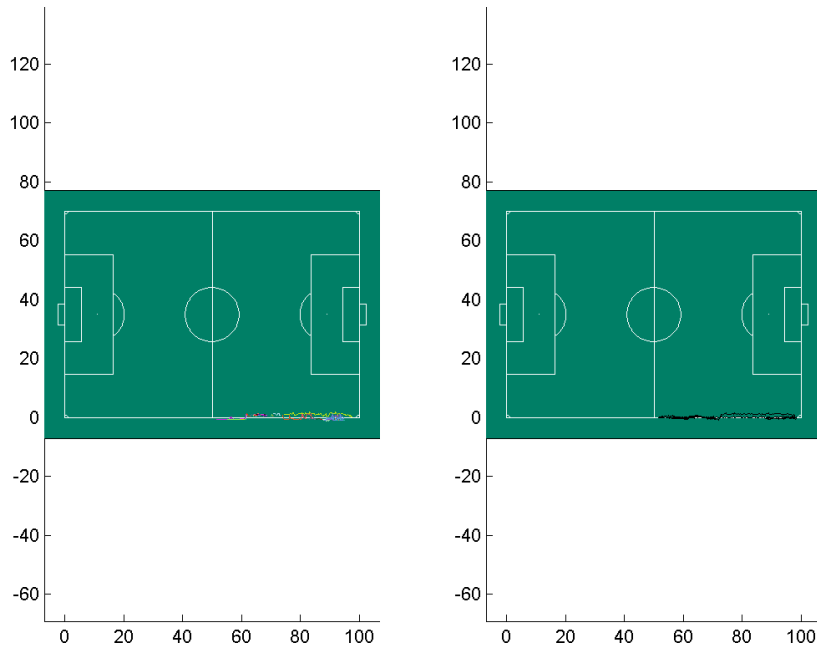


Figure K-2: Supervised associated fragments and complete trajectory for uniqueID = 201

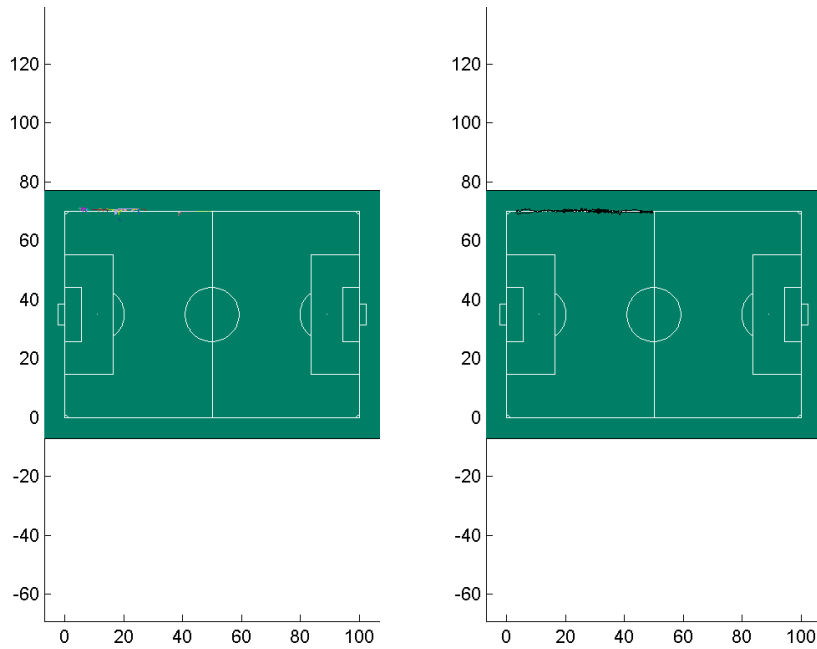


Figure K-3: Supervised associated fragments and complete trajectory for uniqueID = 202

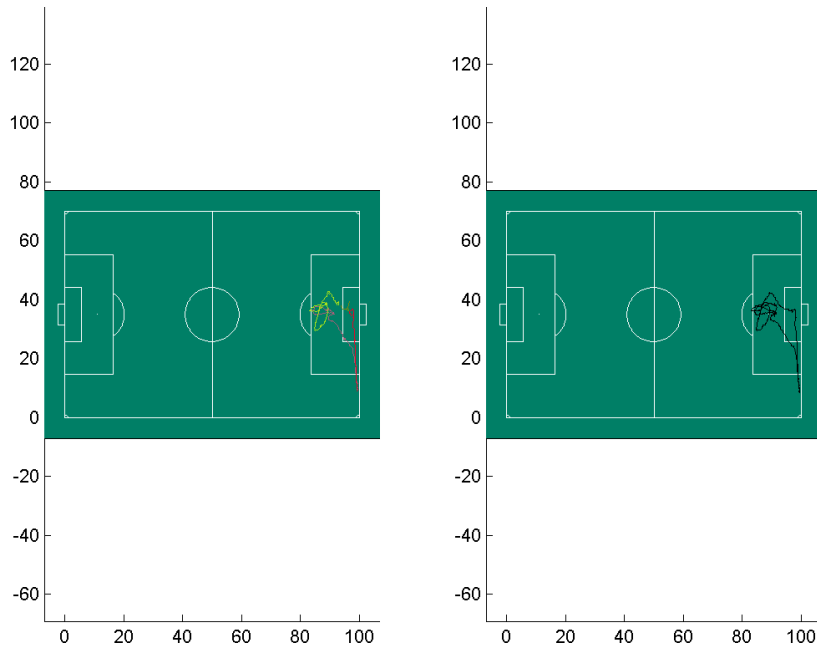


Figure K-4: Supervised associated fragments and complete trajectory for uniqueID = 1

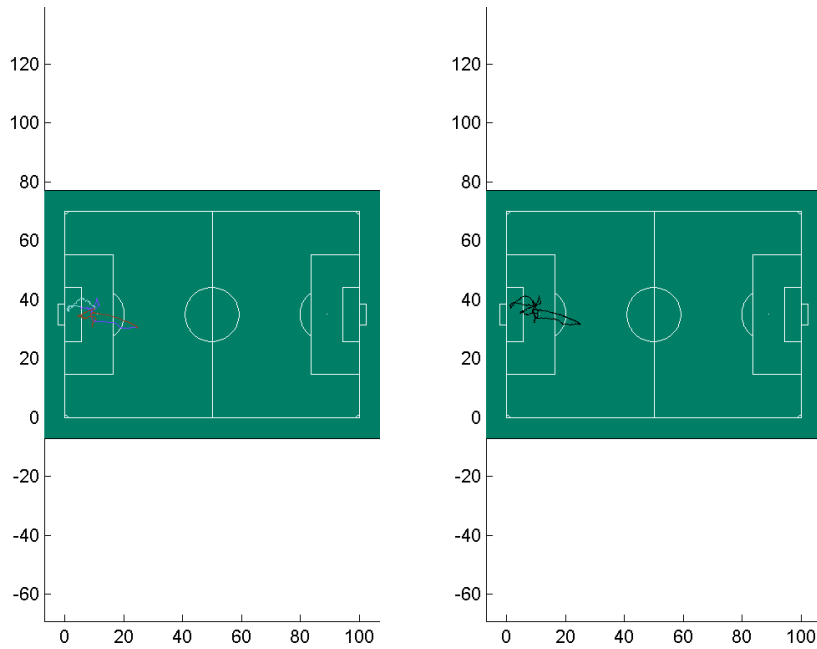


Figure K-5: Supervised associated fragments and complete trajectory for uniqueID = 105

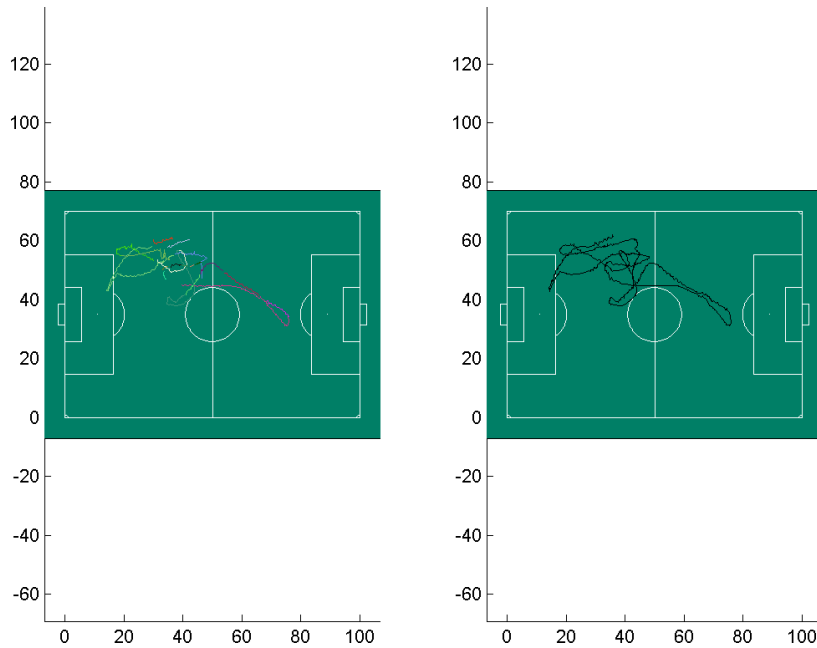


Figure K-6: Supervised associated fragments and complete trajectory for uniqueID = 5

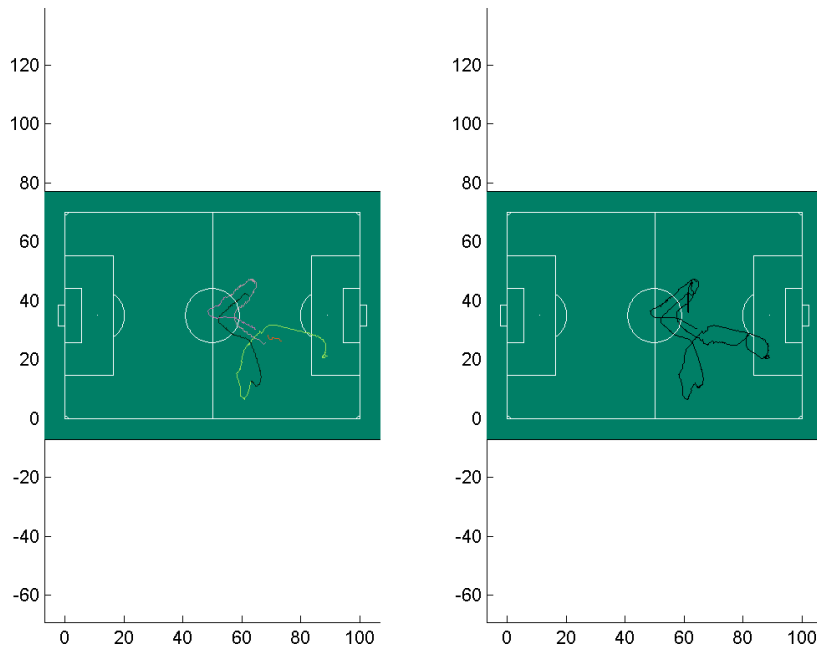


Figure K-7: Supervised associated fragments and complete trajectory for uniqueID = 6

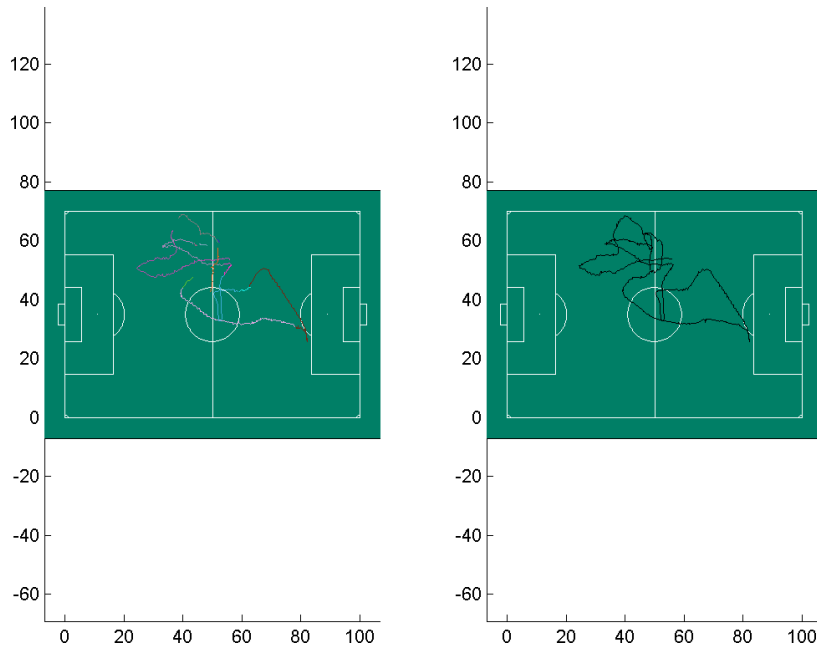


Figure K-8: Supervised associated fragments and complete trajectory for uniqueID = 8

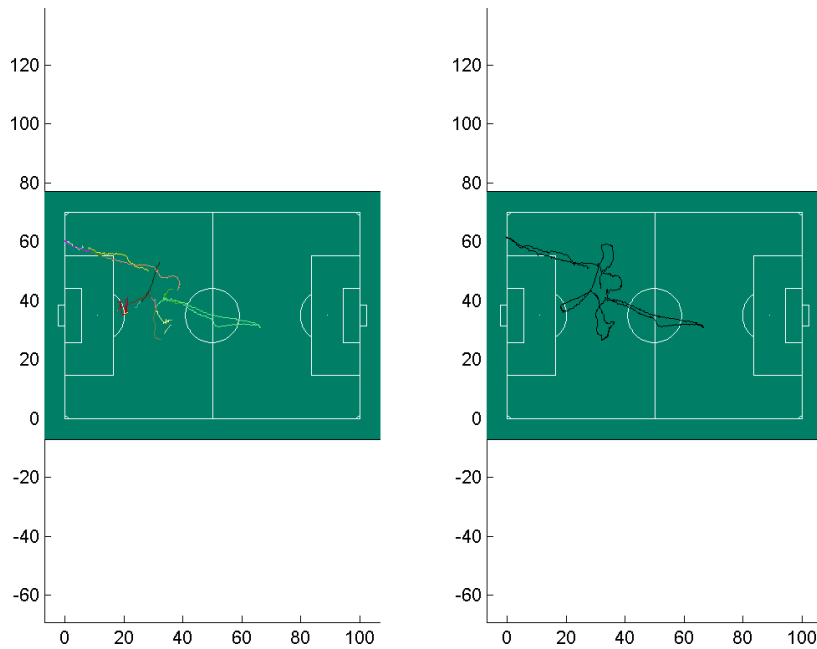


Figure K-9: Supervised associated fragments and complete trajectory for uniqueID = 9

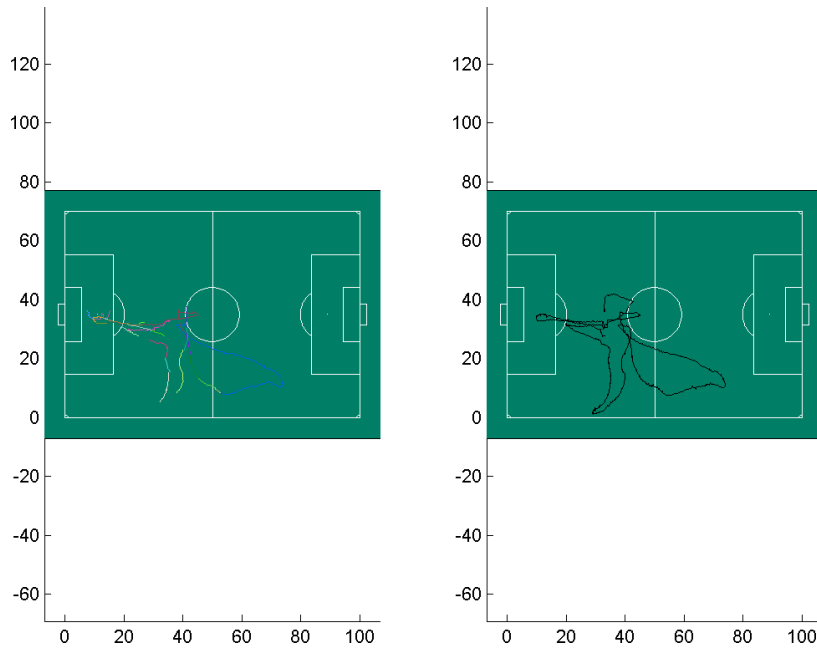


Figure K-10: Supervised associated fragments and complete trajectory for uniqueID = 10

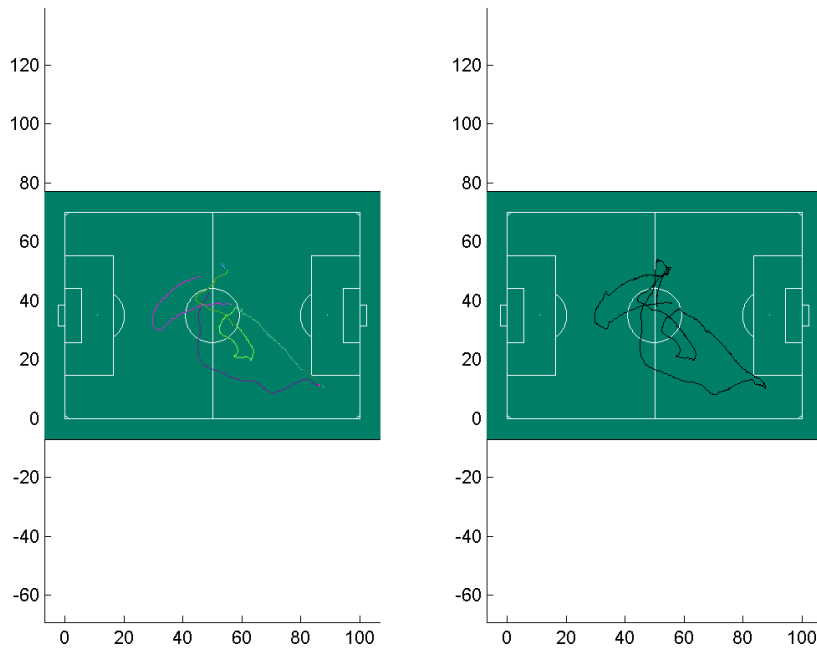


Figure K -11: Supervised associated fragments and complete trajectory for uniqueID = 11

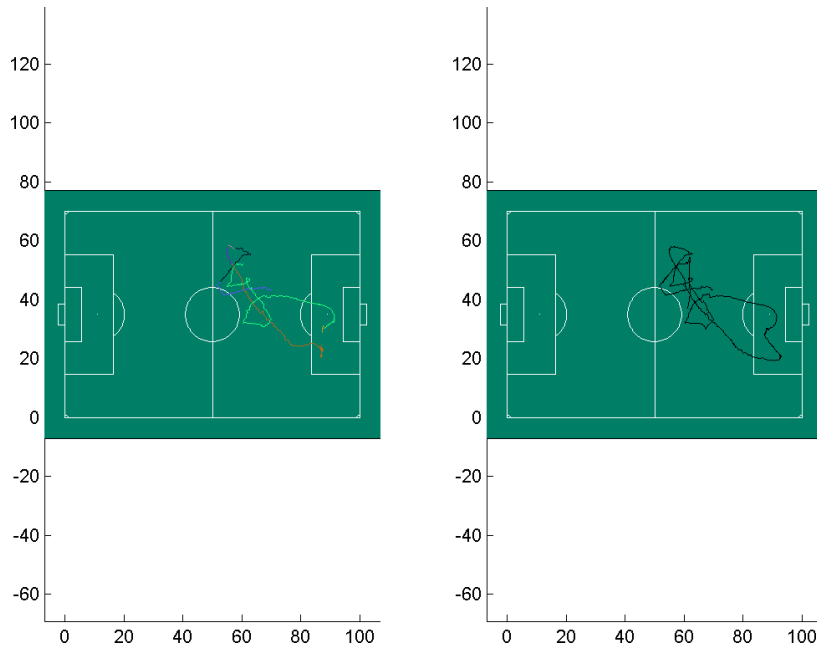


Figure K -12: Supervised associated fragments and complete trajectory for uniqueID = 14

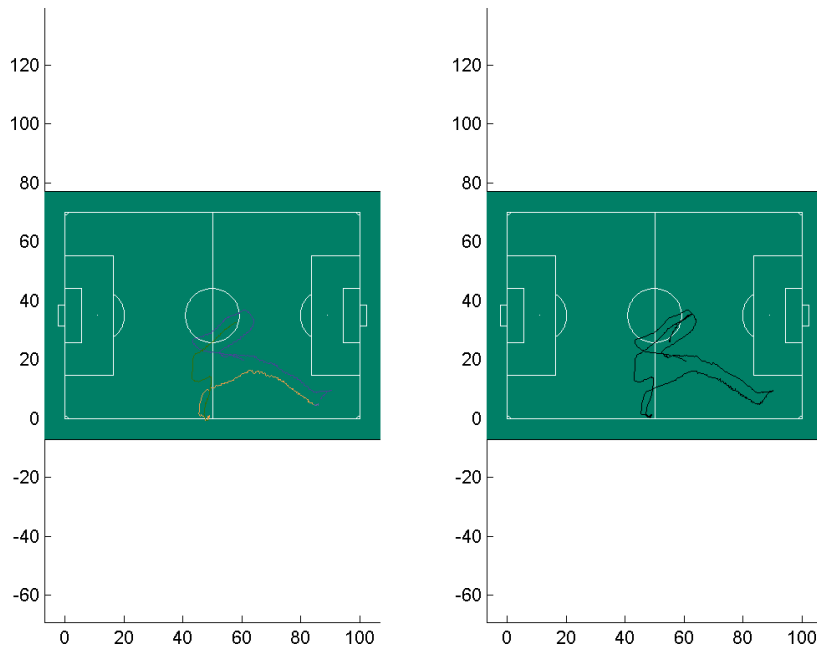


Figure K-13: Supervised associated fragments and complete trajectory for uniqueID = 20

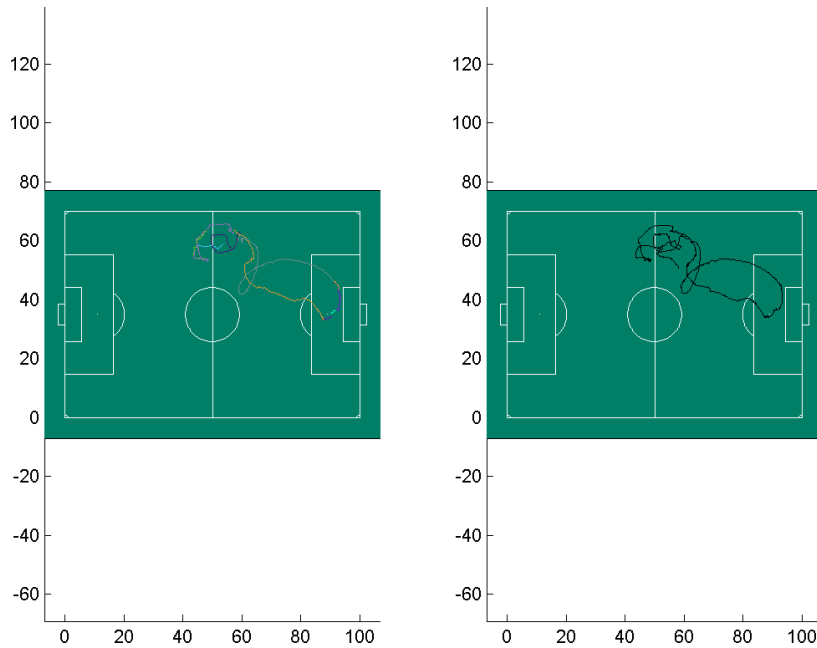


Figure K-14: Supervised associated fragments and complete trajectory for uniqueID = 21

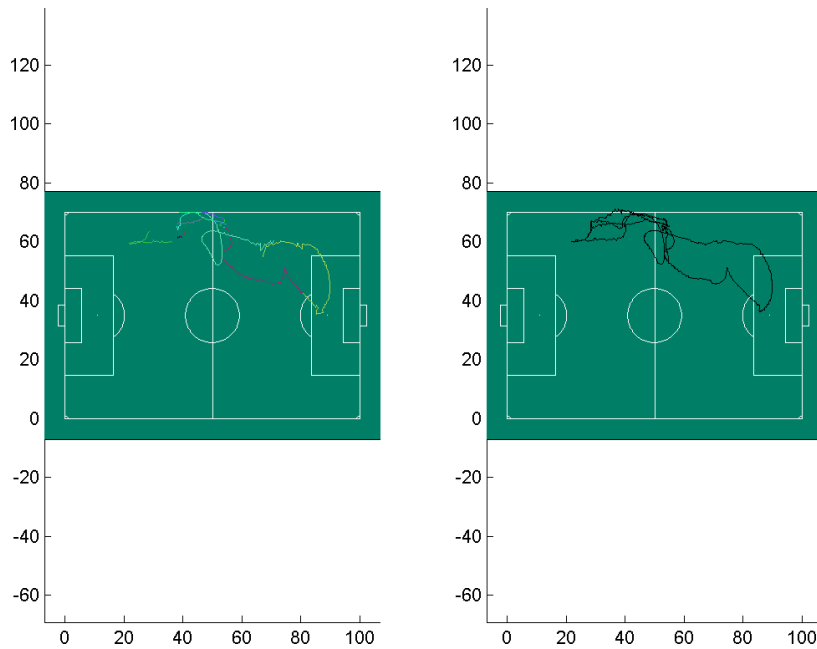


Figure K-15: Supervised associated fragments and complete trajectory for uniqueID = 26

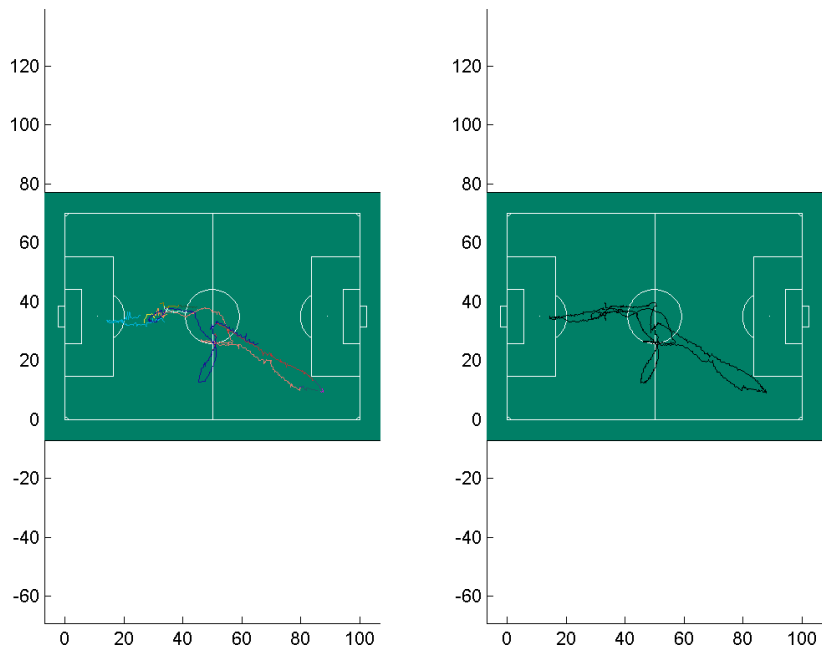


Figure K-16: Supervised associated fragments and complete trajectory for uniqueID = 104

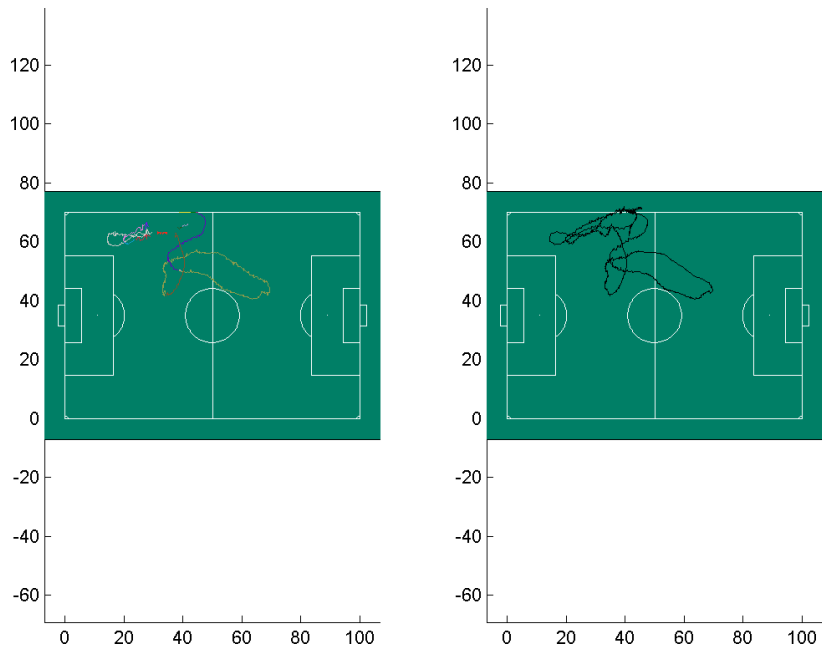


Figure K-17: Supervised associated fragments and complete trajectory for uniqueID = 107

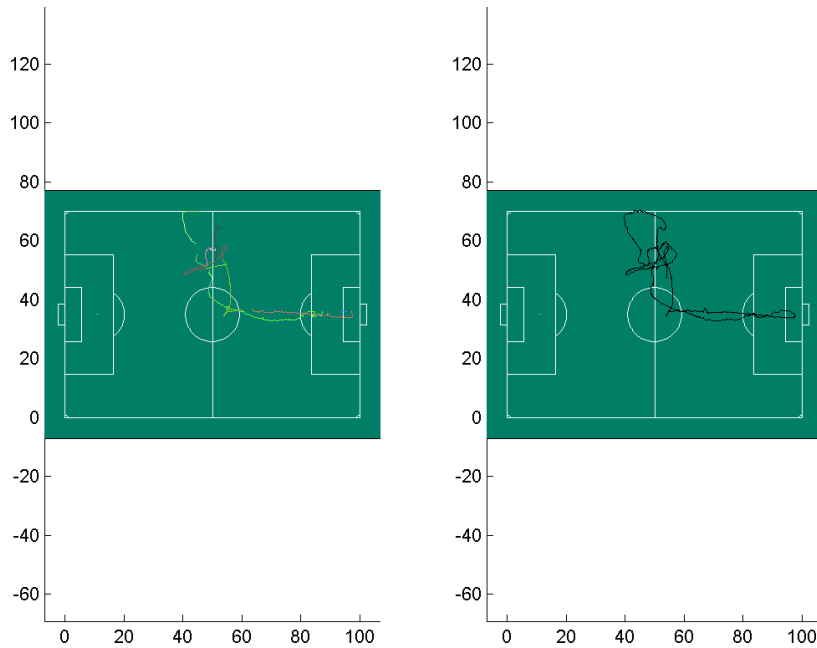


Figure K-18: Supervised associated fragments and complete trajectory for uniqueID = 108

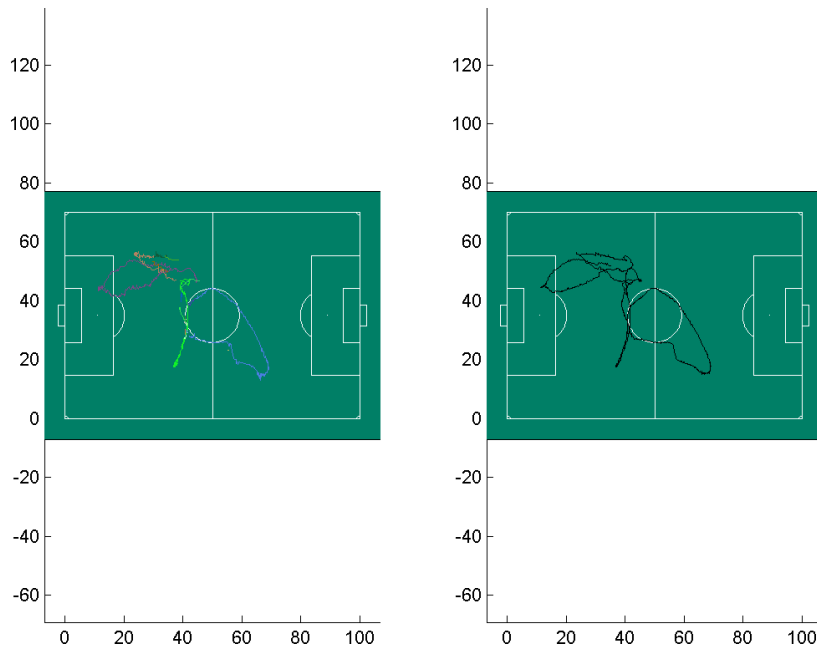


Figure K-19: Supervised associated fragments and complete trajectory for uniqueID = 113

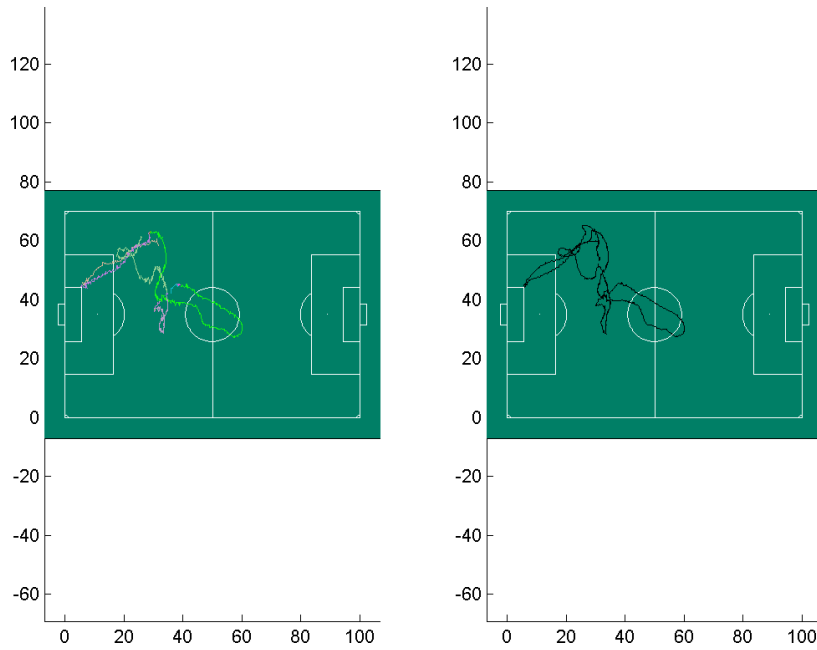


Figure K-20: Supervised associated fragments and complete trajectory for uniqueID = 114

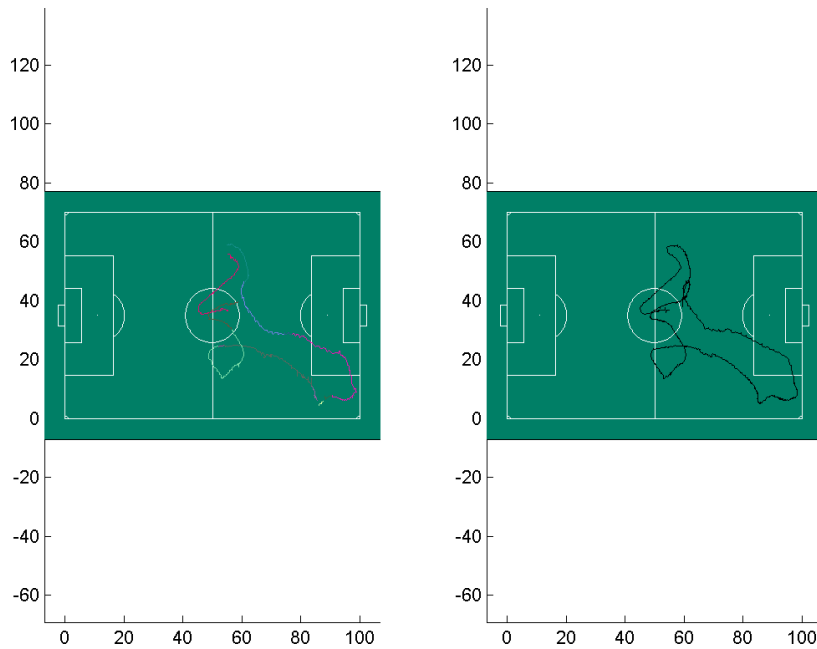


Figure K-21: Supervised associated fragments and complete trajectory for uniqueID = 120

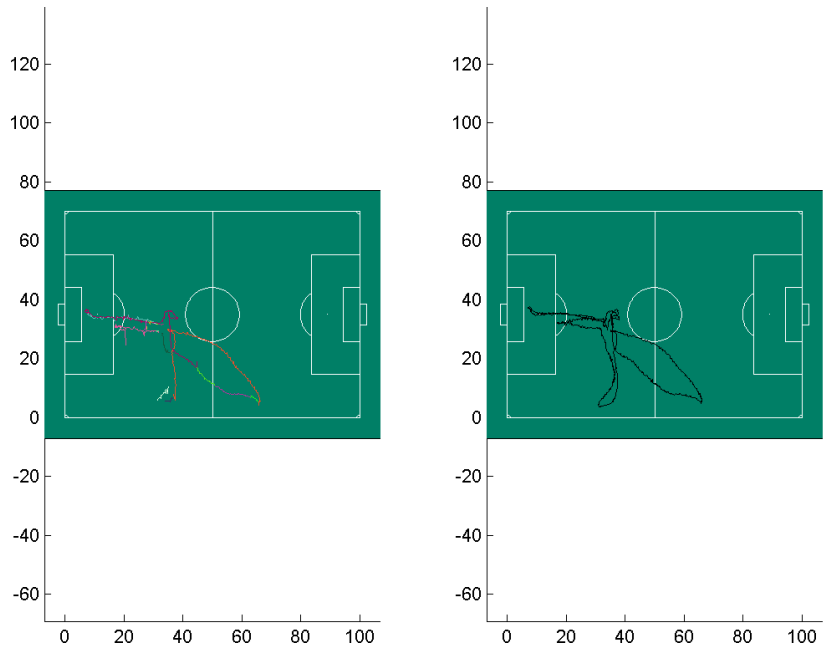


Figure K-22: Supervised associated fragments and complete trajectory for uniqueID = 125

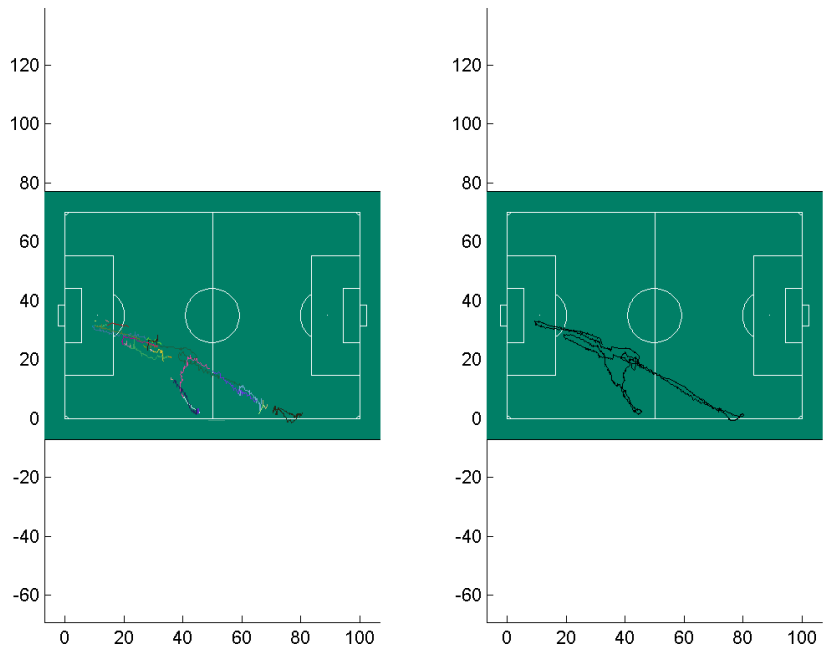


Figure K-23: Supervised associated fragments and complete trajectory for uniqueID = 126

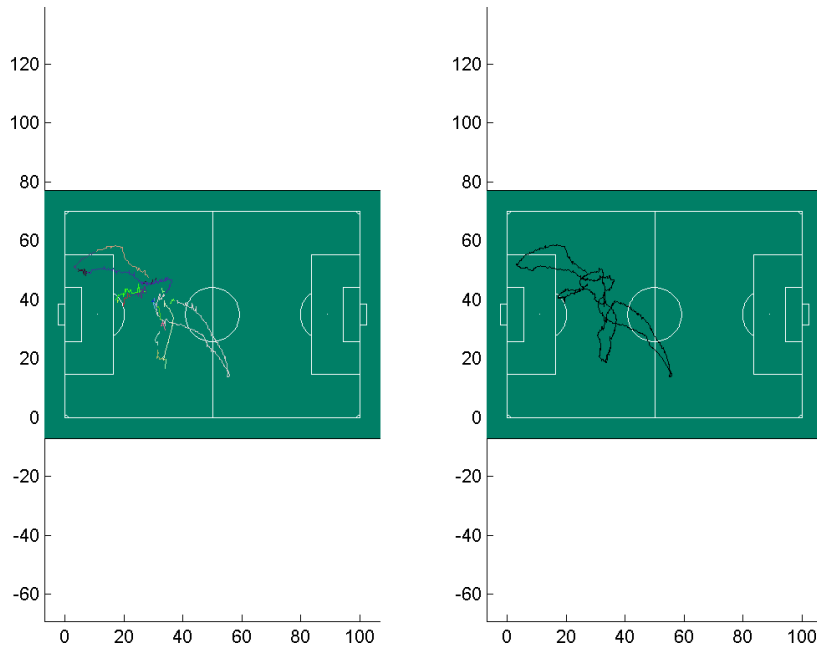


Figure K-24: Supervised associated fragments and complete trajectory for uniqueID = 128

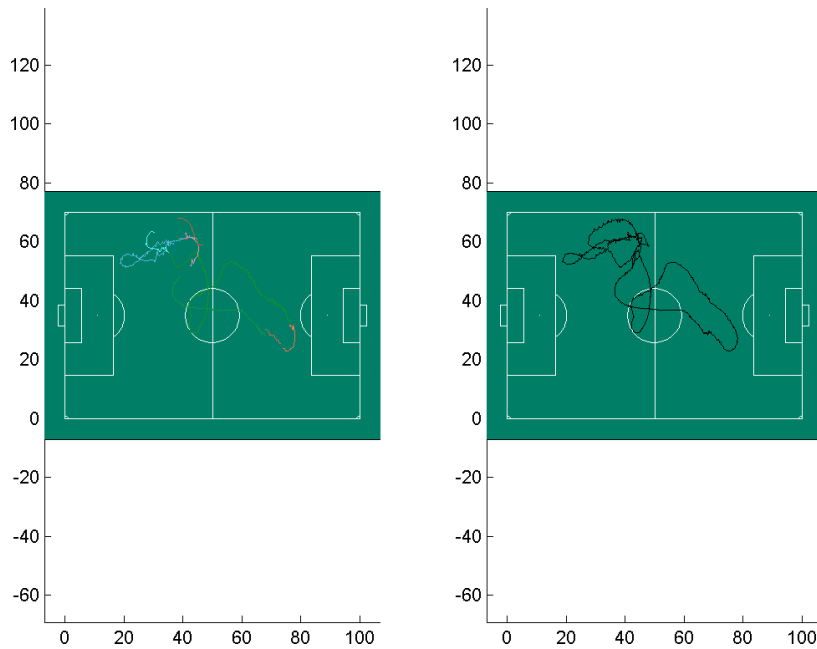


Figure K-25: Supervised associated fragments and complete trajectory for uniqueID = 155

L.Introducción

Motivación

La motivación principal de este proyecto es desarrollar un sistema capaz de detectar y seguir a los jugadores del campo de videos deportivos tomados de distintas posiciones en un sistema multicámara.

Desde una perspectiva de ocio, los videos deportivos representan una proporción significativa del total de emisiones de la televisión pública y comercial. Una gran cantidad de trabajo ya se ha realizado en el análisis de contenido de vídeo de deportes, y el trabajo para mejorar y enriquecer los videos deportivos está creciendo rápidamente debido a la alta demanda de los usuarios.

Desde una perspectiva profesional, el análisis de videos deportivos, especialmente en juegos de balón (fútbol, tenis, baloncesto...) es particularmente útil para analizar y mejorar las tácticas y rendimientos de los jugadores (y equipos). Gracias a estos sistemas, los entrenadores pueden obtener datos, estadísticas y otra información difícil de obtener si un sistema de este tipo. Esto es interesante también para las emisiones deportivas ya que permite obtener estadísticas y hechos interesantes para la audiencia.

Objetivos

El principal objetivo de este proyecto es crear un sistema que, tras una configuración previa (y con cierta supervisión para los deportes de equipo), sea capaz de detectar y seguir a cada jugador del campo. Este sistema debe ser completo, general y modular, facilitando mejoras de trabajo futuro y modificaciones.

El sistema empezará con un sistema base originalmente diseñado para video-vigilancia al que se le añadirán nuevas características y mejoras. Adicionalmente se crearán nuevos módulos necesarios para el correcto funcionamiento del sistema completo.

Los deportes considerados son deportes individuales (ej. tenis) y los deportes de equipo (ej. baloncesto, futbol). Dadas las características de los vídeos mencionados, se diseñará un sistema separado para cada uno de ellos.

Tras obtener los sistemas básicos, se estudiarán y desarrollarán funcionalidades adicionales relacionadas con rendimientos, estadísticas y distintas representaciones de los resultados.

Estructura del documento

La memoria incluye los siguientes capítulos:

- **Capítulo 1.** Motivación y objetivos del proyecto
- **Capítulo 2.** Estado del arte del análisis de videos deportivos y presentación de algunos content sets.
- **Capítulo 3.** Descripción del sistema base y de los módulos comunes, incluyendo las modificaciones del sistema base, las homografías y los scripts para la representación del campo.
- **Capítulo 4.** Trabajo desarrollado para los deportes individuales, describiendo el sistema, la fusión aplicada y los ajustes, pruebas y resultados.
- **Capítulo 5.** Trabajo desarrollado para los deportes de equipo, describiendo el sistema, las distintas fusiones aplicadas, el sistema de evaluación diseñado y los ajustes pruebas y resultados.
- **Capítulo 6.** Estadísticas extraídas de los resultados de los sistemas desarrollados en los capítulos 4 y 5.
- **Capítulo 7.** Conclusiones del proyecto y trabajo futuro.
- **Anexo A.** El sistema base descrito con mayor precisión y detalle.
- **Anexo B.** Teoría matemática de las homografías.
- **Anexo C.** Fondos generados para cada cámara de los content sets usados para futbol y tenis.
- **Anexo D.** Máscaras generadas para el sistema de fútbol.
- **Anexo E.** Representación de los blobs que pertenecen al mismo jugador tras obtener de forma automática el ID único.
- **Anexo F.** Resultados de deportes individuales usando el tracking ideal.
- **Anexo G.** resultados de los deportes de equipo usando el tracking del sistema base.
- **Anexo H.** Comparación entre los resultados del tracking ideal y los resultados del tracking del sistema base para deportes de equipo.

- **Anexo I.** Estadísticas de los jugadores extraídas de los resultados del sistema base.
- **Anexo J.** Figuras de las trayectorias resultantes del sistema de futbol.
- **Anexo K.** Asociación supervisada de los fragmentos de las trayectorias de cada jugador para el sistema de futbol.

M. Conclusiones

El objetivo principal de este proyecto, diseñar y desarrollar un sistema para detectar y seguir jugadores en un campo usando videos multicámara, ha sido conseguido. Tras una configuración previa y con cierta supervisión (para los deportes de equipo), el sistema es capaz de detectar y seguir a cada jugador en el campo, y de proporcionar algunas estadísticas de los jugadores. El sistema es completo, general y modular, facilitando mejoras y modificaciones futuras.

Los videos deportivos con cámaras fijas tienen características importantes que ofrecen ventajas para su análisis con respecto a los videos generales de video-vigilancia. Los fondos son generalmente estáticos y uniformes, excepto en ciertas zonas como el público o los anuncios dinámicos, que pueden modelarse de forma relativamente simple. Las personas seguidas tienen uniforme específico y distinto, al menos entre los distintos tipos de personas (jugadores de cada equipo, porteros, árbitros...). Esta última característica es también una desventaja ya que los jugadores de un mismo equipo tienen exactamente la misma apariencia ya que llevan el mismo uniforme, lo que complica el seguimiento cuando se producen oclusiones.

La colocación de las cámaras es importante. El caso ideal es cuando las cámaras están enfrentadas o situadas simétricamente, ya que en esos casos los errores de tracking se reducen en el proceso de fusión.

Los casos con mayor solapamiento entre cámaras presentan mejores resultados. Esta característica se observa en el caso de deportes de equipo, cuando se compara el resultado de las cámaras enfrentadas con el resultado de la fusión de regiones.

La colocación de las cámaras a mayor altura es interesante ya que reduce el error de seguimiento. A mayor altura de la cámara, menor área de píxeles se proyecta en el plano del campo y por lo tanto se asegura una mayor precisión en la proyección de la homografía, pero si la cámara está situada a demasiada altura, la identificación del uniforme del jugador puede ser complicada. Además, usar el punto medio de la base reduce el error ya que, como se explica en la sección correspondiente, se reduce la distancia entre el centro del jugador y el punto escogido.

En el caso de deportes individuales el sistema es relativamente simple. Si el fondo se mantiene estático y el seguimiento funciona relativamente bien, el resultado es apropiado y

cumple los objetivos. Al no haber oclusiones se facilita el seguimiento y la determinación de a quien pertenece cada blob.

En el caso de deportes de equipo, los problemas principales son las oclusiones y las zonas con bajo o nulo solapamiento de cámaras. Estos sistemas son los que necesitan mas mejoras, como se ve en el estado del arte y en los resultados experimentales. Hay muchos métodos de fusión y muchos parámetros que cuando se combinan adecuadamente contribuyen a mejores resultados. En las mejoras hechas se han conseguido mejoras significativas con cambios relativamente sencillos. Como se muestra en los experimentos, en el caso del sistema real de tracking es necesario un post-procesado supervisado que permite obtener las trayectorias completas, y cuya sustitución por un proceso automático se propone como trabajo futuro.

N. Publicaciones generadas

Confirmation mail

Fecha: 29 Aug 2012 12:03:41 -0400 [18:03:41 CEST]

De: [Multimedia Tools and Applications <angie.malanday@springer.com>](mailto:angie.malanday@springer.com)

Para: [Rafael Martín <rafael.martinn@estudiante.uam.es>](mailto:rafael.martinn@estudiante.uam.es)

Asunto: New Submission

Dear Rafael Martín:

Thank you for submitting your manuscript, "A semi-supervised system for players detection and tracking in multicamera soccer videos", to MULTIMEDIA TOOLS AND APPLICATIONS.

During the review process, you can track the status of your manuscript by accessing the following web site:

<http://mtap.edmgr.com/>

Your username is: RafaelMartin

Your password is: *****

With kind regards,

The Editorial Office
MULTIMEDIA TOOLS AND APPLICATIONS

Multimedia Tools and Applications

A semi-supervised system for players detection and tracking in multicamera soccer videos

--Manuscript Draft--

Manuscript Number:	
Full Title:	A semi-supervised system for players detection and tracking in multicamera soccer videos
Article Type:	Manuscript
Keywords:	sports videos; multicamera systems; object detection; object tracking; fusion; homography
Corresponding Author:	Rafael Martín Colmenar Viejo, SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Rafael Martín
First Author Secondary Information:	
Order of Authors:	Rafael Martín José María Martínez, PhD
Order of Authors Secondary Information:	

A semi-supervised system for players detection and tracking in multicamera soccer videos

Rafael Martín, José M. Martínez

VPULab, EPS – Universidad Autónoma de Madrid, Spain

rafael.martinn@estudiante.uam.es, josem.martinez@uam.es

Abstract This paper presents a system which after a simple previous configuration is able to detect and track each player on the court or field. This system is complete, general and modular, to be improved and modified by future work. The system is based on a base system, originally designed for video surveillance, which has been modified for the team sports domain. Additionally, the necessary modules for the proper functioning of the full system have been developed. Target sports of the developed system are team sports (e.g., basketball, soccer). In addition to the detection and tracking system, an evaluation system is designed which allows to obtain quantitative results of the system performance.

Keywords *sports videos, multicamera systems, object detection, object tracking, fusion, homography.*

1 Introduction

Sports broadcasts constitute a major percentage in the total of public and commercial television broadcasts. A lot of work has already been carried out on content analysis of sports videos, and the work on enhancement and enrichment of sports video is growing quickly due to the great demands of customers.

An overview of sports video research is given in [1], describing both basic algorithmic techniques and applications. Sports video research can be classified into the following two main goals: indexing and retrieval systems (based on high-level semantic queries) and augmented reality presentation (to present additional information and to provide new viewing experience to the users). The analysis of events can be carried out by combining various attributes of the video, including its structure, events and other content properties. In event detection, features can be extracted from three channels: video, audio and text. The definition of what is interesting differs for each viewer. While sport fans are more interested in events like goals or spectacular points in tennis, the coaches might be more interested in the errors of the players to help them improve their capabilities.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

In addition, algorithmic building blocks must consider structure analysis (for example of a match), object detection and segmentation (e.g., position of players, shape and movements of the athletes), ball and player tracking (specially the tracking of a small ball is a difficult problem), camera calibration (to determine the geometric mapping between the image coordinates and the real world positions and 3D reconstruction: multiple cameras can be used to obtain views from different position and compute a complete reconstruction of the playfield).

The main applications of sports video processing are:

- *Abstracting*: creating a shortened version of a video that still comprises the essential information from the complete video.
- *Tactic and performance analysis*: to understand the tactics that teams or individual players have used and to evaluate the performance of a team or a player through analyzing their motion and activity in games.
- *Augmented reality presentation of sports*: with two basic techniques: 3D reconstruction of the game to provide arbitrary views, and to insert some illustrations into the original video or to provide the illustration in extra windows to help consumers to understand the video easier.
- *Sport video for small devices*: the transmission of sport events on small devices like PDA, 3G/UMTS phones will become more popular in the future
- *Referee assistance*: to help the referee in cases of difficult decisions and partially replace the work of the referee.

There are two main types of sports video, used to perform the analysis: edited broadcast and multi-camera.

- *Edited broadcast* [2][3][4][5]: videos are edited for broadcasting on television. On this type of videos, there are scenes of the game, repetitions and changes of viewpoint of the observer. This type of videos is the most common of all.
- *Multi-camera*: In general, multi-camera videos are not edited. They are subdivided into two main classes:
 - *Mobile* [6]: Cameramen follow the players, changing the orientation of their cameras. The obtained videos usually are used to generate the broadcast videos, mixing and editing parts of them.

- *Fixed* [7][8]: The cameras remain fixed from the beginning to the end of the recording. In this case, no operator is needed for controlling the orientation of the camera.

The presented system is centered on team sports. We refer team sports for sports where two or more players move in the same areas of the field. Referees may also be moving in the same areas and among the players.

For team sports, fusion is harder than for individual sports, because all the tracked blobs do not belong to the same player. During tracking, occlusions may occur, causing problems in tracking as losing players or generating fused blobs from more than one player.

Some examples of this kind of sports are soccer, basketball or volleyball.

The remainder of the paper is structured as follows: after this introduction first section, section 2 presents a brief overview of the related work and the used dataset; Section 3 describes the designed system; Section 4 explains the fusion method; Section 5 depicts the evaluation system; Section 6 presents the adjustments, testings and results of the implementation of the system; Section 7 show some applications of the system results; Finally, section 8 describe the conclusions and the future work.

2 State of the art

2.1 Sport video analysis techniques

The canonical system for detecting and tracking players in a field can be decomposed into several main blocks, as depicted in Figure 1. The different techniques and associated algorithms that are key for any sports video analysis system are described below.

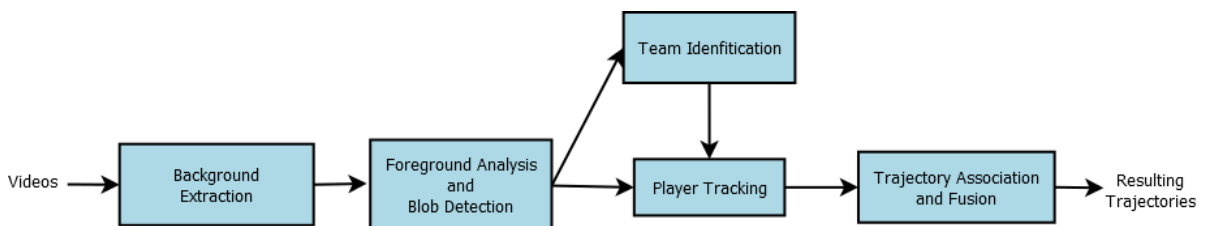


Fig. 1: Block diagram of the canonical system

Background extraction is a simple and very common method used for moving objects segmentation which consists of the difference between a set of images and the background model. The background model is generated from an empty image of the field (if it is available) or from a fragment of the video with non-static players. Some of the main problems in background extraction are the changes in the environment such as illumination, shadows

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

or background moving objects. Some statistical adaptive methods are used in [2] for this block. A histogram learning technique is employed in [3] to detect the playfield pixels, using a training set of soccer videos. An example of the playfield detection result is shown in Figure 2



Fig. 2: Playfield detection result: Original frame (left) and the detected playfield pixels (right)

In [5], background is modeled with a mixture of Gaussians and learned before a match without the need of an empty scene. Background detector is first used to extract a pitch mask for saving processing time and avoiding false alarms from the crowd. A background model is generated in [6], and a background updating procedure is used to continuously adapt the model to the variations in light condition. In [7], the region of the ground field is extracted, first by segmenting out non-field regions like advertisements, and then the players are extracted on this field on the basis of the field extraction. The field is represented in [8] by a constant mean color value that is obtained through prior statistics over a large data set. The color distance between a pixel and the mean value of the field is used to determine whether it belongs to field or not. In the system, only hue and saturation components in HSV color space are computed to exclude effect of illumination.

Blob detection is based on a *foreground analysis* to identify the players. The goal is to adjust as much as possible to the contour of the player blob, separating the near or overlapping possible players. In [2], the nodes are grouped considering the edges between them. The area of the blobs is the parameter used to define the number of components of each node. The graph is constructed from the set of blobs obtained during the segmentation step in such a way that nodes represent blobs, and edges represent the distance between these blobs. In [3], connected component analysis (CCA) scans the binary mask and groups its pixels into components based on pixel connectivity. The foreground mask is morphologically filtered in [4] (closing after opening) to get rid of outliers. This is followed by a connected component analysis that allows creating individual blobs and bounding boxes are created. In [5], for an isolated player, the image measurement comes from the bottom of foreground region directly. The measurement covariance in an image plane is assumed to

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

be a constant and diagonal matrix, because foreground detection in an image is a pixel-wise operation. A further step with the connectivity analysis has been introduced in [6] to extract connected regions and remove small regions due to noise. This procedure scans the entire image and groups neighbor pixels into regions. The connectivity analysis eliminates shadows by using geometrical considerations about the shape of each region.

The information of the uniform of the teams is used for the *team identification* of the blob and difference between players, goalkeepers or referees. Generally a player can be modeled as a group of many regions, each region having some predominant colors.

Dividing the model of the player in two or more regions is attempted in [2], so that each region represents a part of the team's uniform, e.g, t-shirt, short, socks. For each region, a filtering based on the vertical intensity distribution of the blobs is defined. An example of vertical intensity distribution is shown in Figure 3.

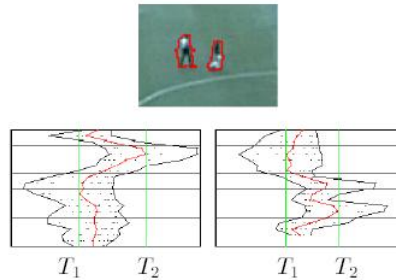
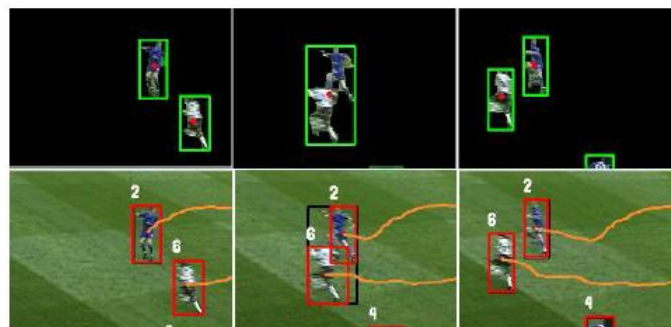


Fig. 3: Example of vertical intensity distribution

Color histograms are used to classify the objects of templates for each class in [4]. The different classes consist of team 1, team 2, goalkeeper team 1, goalkeeper team 2, and referee. The center part of the detected blobs is taken and color histograms are calculated. Histograms are compared, after normalization, with the color histograms of the templates using the Bhattacharyya distance. This block can be also implemented using a histogram-intersection method[5]. The result for each player is a five-element vector, indicating the probability that the observed player is wearing one of the five categories of uniform (the categories are the same as described in [4]). In [7], the group behavior of soccer players is analyzed using color histogram back-projection to isolate players on each team. But since different teams can have a similar histogram, vertical distribution of colors is used.

When the position of each player position is detected, the next objective is to get the *player tracking* to know where the player is at each moment. In [2], the tracking of each player is performed by searching an optimal path in the graph defined in blob detection. At each step, the minimal path in the graph is considered, using the distance information between

1 the blobs. The bounding box and centroid coordinates of each player are used in [5] as
2 state and measurement variables in a Kalman filter. Finally, in track selection, there is a
3 procedure of tracking aided recognition for the 25 most likely players (11 from each team
4 and 3 referees). Due to false alarms and tracking errors, there normally exist more than 25
5 tracks for the players. A player likelihood measure is calculated for each target on the basis
6 of confidence of category estimate, number of support cameras, domain knowledge in po-
7 sitions (for goalkeepers and linesmen), frames of being tracked or missing, as well as the
8 fixed population constraint. A fast sub-optimal search method gives reasonable results. An
9 example of player detection and tracking is shown in Figure 4.
10
11
12
13
14
15

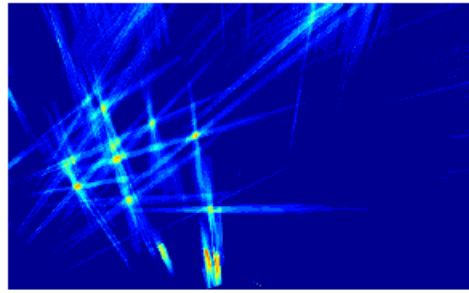


16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
Fig. 4: Player detection and tracking from a single camera

The multiplayer tracker may use the Maximum a Posteriori Probability[6] to get the best fit of state and observation sequences. The state vector includes information about the location, velocity, acceleration and dimension of the single bounding box. Template matching and Kalman filter is used in [7]. The templates of players are extracted from player mask using connected component extraction. First, new players that do not significantly overlap with the bounding box of a player already tracked are found out. Then, the new players are inserted to tracking list. Location of players at the next frame is predicted by Kalman filter and template matching at that location is performed. Finally, the player template is updated. The main problem of player tracking is occlusion and in [7] only occlusion between different teams is considered. In [9], starting from a frame where no two players occlude each other, players are tracked automatically by considering their 3D world coordinates. When a merging of two (or more) players is detected, separation is done by means of a correlation-based template matching algorithm.

The *trajectory association* across the multiple points of view requires the *fusion* of multiple simultaneous views of the same object as well as the fusion of trajectory fragments captured by individual sensors. There are two main approaches for generating the fused trajectories: fuse and then track [10], and track and then fuse [5][6][11][12][13].

1 A multi-camera multi-target track-before-detect (TBD) particle filter that uses mean-shift
2 clustering is presented in [10]. A sample projection of detection mask is shown in Figure 5.



14 Fig. 5: sample projection of detection mask from multiple cameras to a top view

16 The information from multiple cameras is first fused to obtain a detection volume using a
17 multi-layer homography. To track multiple objects in the detection volume, unlike tradi-
18 tional tracking algorithms that incorporate the measurement at the likelihood level, TBD
19 uses the signal intensity in the state representation. The association matrix also can be de-
20 cided according to the Mahalanobis distance[5] between the measurement coordinates and
21 the track prediction. To solve the correspondence problem, a graph based algorithm can be
22 applied [6]. The best set of tracks is computed by finding the maximum weight path. This
23 step can be performed using the algorithm by Hopcroft and Karp. In [11], after obtaining
24 the transformed trajectories, the next step is to compute their relative pair-wise similarities
25 for association and fusion in order to have a single trajectory corresponding to an object
26 across the entire field. The parameters used for the association in [12] are: shape, length,
27 average target velocity, sharpness of turns (which defines the statistical directional charac-
28 teristics of a trajectory), trajectory mean and the PCA component analysis. With all those
29 parameters the parameter vector is generated and the cross correlation is used as proximity
30 measure. Multi camera analysis algorithms that generate global trajectories may be sepa-
31 rated in three main classes, namely: appearance based[7][14][15], geometry
32 based[16][17][18][19] and hybrid approaches[13][20][21] [22]. Following[11], the main
33 ideas of each class are described below. Appearance-based approaches use color to match
34 objects across cameras. Geometry-based approaches establish correspondences between
35 objects appearing simultaneously in different views. Hybrid methods use multiple features
36 to integrate the information in the camera network.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

The following table presents an overview of all the reviewed sport video analysis techniques:

Reference	Sport	Background extraction	Foreground analysis and blob detection	Team identification	Player tracking	Trajectory association and fusion
[2]	Football	Statistical adaptive methods, Gaussian distribution	Graph representation	Vertical intensity distribution	Optimal path in the graph	-
[3]	Football	Learning histogram	Connected component analysis	-	-	-
[4]	Football	-	Morphological filters, connected component analysis	Color histogram contrast	-	-
[5]	Football	Mixture of Gaussians	Measurement covariance	Histogram intersection	Kalman filter	Mahalanobis distance
[6]	Football	Background with updating procedure	Connectivity analysis	-	Maximum a Posteriori Probability	Graph based algorithm, Hopcroft and Karp algorithm
[7]	Football	Peak values of histogram, morphological filtering	-	Vertical distribution of colors	Template matching, Kalman filter	-
[8]	Football	Mean color, color distance	-	-	-	-
[9]	Football	-	-	-	Template matching	-
[10]	Basketball	-	-	-	-	Particle filter, multi-layer homography, K-means, mixture of Gaussian clustering
[11]	Football	-	-	-	-	Feature vector correlation
[12]	Basketball and football	-	-	-	-	Feature vector correlation
[13]	-	-	-	-	-	Graph based algorithm

Table 1: Overview of sport video analysis techniques

2.2 Content set: ISSIA Soccer Dataset[23]

The public ISSIA dataset¹ is composed of:

- Six synchronized views acquired by six Full-HD cameras, three for each major side of the playing-field, at 25 fps (6 AVI files).
- Manually annotated objects position of two minutes of the game. These metadata provide the positions of the players, referees, and ball in each frame of each camera (6 XML files). The players have the same labels while they move in the six views that correspond to the numbers on their uniforms. The player labels of the first team start from 1 while the player labels of the second teams start from 201.

The first 300 frames of each sequence have not been labeled in order to provide an initial phase to initialize the background subtraction algorithms.

- Calibration data: Pictures containing some reference point in to the playing- field and the relative measures for calibrating each camera into a common world coordinate system (6 pdf files).

All cameras are DALSA 25-2M30 cameras.

The positions of the six cameras on the two sides of the field are shown in Figure 6.

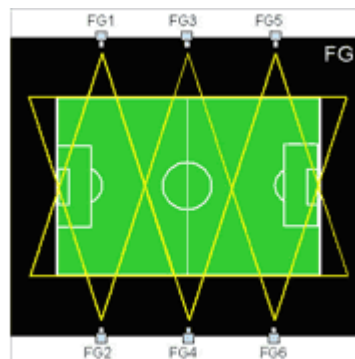


Fig. 6: Positions of the six cameras of ISSIA Soccer Dataset

¹ <http://www.issia.cnr.it/htdocs%20nuovo/progetti/bari/soccerdataset.html>

3 System

In this section the global system is presented. The system block diagram is depicted in Figure 7.

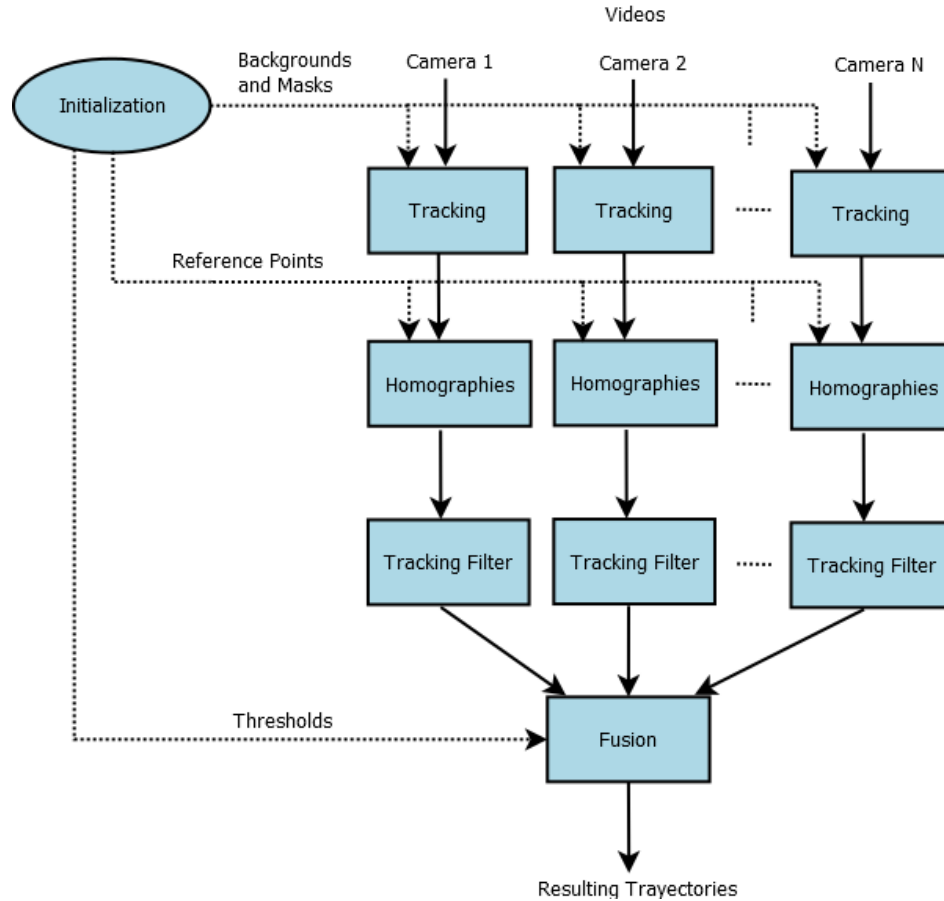


Fig. 7: System block diagram for team sports

The **Initialization block** is executed first. It generates the field representation, the backgrounds, the masks, the homography reference points and the thresholds for the fusion. The Thresholds for fusion are defined in the following section.

The **Tracking blocks** receive the videos and the backgrounds and masks of the different cameras which are recording the players and generate the tracking data for the different perspectives.

Each **Homographies block** receives the tracking of the players from each camera and the reference points, and it generates the top view tracking from each camera.

The **Tracking filter block** receives and filters the tracking for the top view tracking from each camera. Some examples of filtering criteria are: possible regions in the field, minimum number of frames for the blob, maximum width or height for the blob. The block generates the top view filtered tracking from each camera.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Finally, the *Fusion block* receives the thresholds, which are defined in the initialization block, joins the processed tracking from each camera and generates the overall trajectories along the entire field. The fusion method will be described in the next section.

4 Fusion

For the fusion process, the following information is obtained for each blob: number of frames, initial frame, final frame, number of frames and coordinates (x and y) for each frame.

A threshold is defined to decide which blobs should be fused. All the scores under the threshold indicate that both blobs belong to the same player. The three methods to calculate the score are explained in the following subsections. Each score is calculated from two blobs, one from each tracking². Each one of the two blobs belongs to a different tracking. The two trackings must have an overlapping field zone for the fusion. This threshold can be calculated from a training sequence with the evaluation system presented in section 5.

When a score is obtained for each pair of blobs from two different trackings, the next step is to obtain a list of blobs associations (LOA) containing one row for each blob of one of the trackings. Each row indicates which blobs in the other tracking should be fused with the blob that corresponds to that row. A similar LOA for the other tracking is not necessary to calculate because it is redundant, although sometimes it is useful for simplifying further processing (e.g. in the step where the resulting blobs of the fusion process are calculated).

Figure 8 shows an example a LOA.

	1	2
1	5	15
2	20	0
3	0	0
4	21	0
5	4	19
6	0	0

Fig. 8: Example of LOA

The information extracted from the figure is: blob 1 of the first tracking is fused with blobs 5 and 15 of the second tracking (if there were more blobs fused, more columns will appear). Blob 2 of the first tracking is fused with the blob number 20 of the second tracking. Blob 4 of the first tracking is fused with the blob number 21 of the second tracking. Blob 5

² In this case the tracking is defined as the set of blobs that is fused with another set of blobs

of the first tracking is fused with blobs 4 and 19 of the second tracking. Blobs 3 and 6 do not fuse with any blob.

Finally, for the fusion, the LOA is processed to get the group of blobs in each camera that should be fused.

The three proposed fusions have been developed through an incremental development. Each improvement is proposed to solve problems found in the previous fusion. In the tests (see section 6), the results of the three fusions are shown to compare the performance of each fusion.

In the following three subsections each of the three fusions developed are described. The first of the fusions uses only information of the position of each player. The second fusion adds to the first fusion color information of the different uniforms to discriminate between players. The third fusion improves the previous two fusions, adjusting spatially the projection of the trajectories of each player.

4.1 Basic fusion

Given two blobs, the score is defined as the mean square distance per frame, for those frames in which there are both blobs.

$$\begin{aligned} \text{Score}(\text{blobI}, \text{blobJ}) &= \\ &= \begin{cases} \sum_{f=n}^m ((x_{f,i} - x_{f,j})^2 + (y_{f,i} - y_{f,j})^2), & \text{If there are frames in which both blobs appear} \\ \infty, & \text{Else} \end{cases} \end{aligned}$$

Where $x_{f,i}$ and $y_{f,i}$ are the coordinates of the blob I in the frame f , $x_{f,j}$ and $y_{f,j}$ are the coordinates of the blob J in the frame f , and $n \dots m$ are the frames where blob I and blob J are tracked simultaneously.

4.2 Fusion using color information

This type of fusion is the first improvement introduced. The condition when score is not equal to infinity (“if there are frames in which both blobs appear”) is changed to: If there are frames in which both blobs appear and if both blobs wear the same uniform.

This improvement aims increasing the precision parameter, because false fusions are reduced.

4.3 Fusion adjusting correspondence between homographies

Sometimes, resulting trajectories from the two trackings does not fit correctly. This malfunction can be due to imperfections of the camera lens, different camera orientation and height, etc. This problem is observed in the trajectories of the left side of the Figure 9.

In these cases, after seeing the resulting trajectories of the homography projection (for example with a training sequence), the points that define the homography can be changed to obtain a better fit.

Another solution that allows more freedom is to apply a correction to the trajectories after applying the homography. In this way additional problems can be corrected, compared to those obtained simply by changing the points of the homography. Some examples of these additional problems are barrel distortion and pincushion distortion.

Figure 9 and 10 show how this improvement results in a better fit between trajectories of blobs from the different trackings.

5 Evaluation system

For the evaluation system, in addition to the data described in the fusion section, the unique ID of the blob is needed, corresponding to the player or referee. This ID is obtained from the ground truth³. This ID allows knowing if two blobs belong to the same player and evaluating how many fusions are correct or erroneous.

Two blobs should be fused if they belong to the same player (both have the same unique ID) and if there are frames in which both blobs appear. For the fusion of two trackings, a list like the LOA described in section 4 (see Fig 5-2) is obtained but with the certainty that in this case the ideal list of blobs associations (iLOA⁴) is obtained, with all the correct fusions. This iLOA allows calculating the Precision and Recall (defined at the end of this section) when it is compared with the experimental list of blobs associations (eLOA).

After obtaining the iLOA and defining a set of values for the threshold defined in Section 4, eLOAs are obtained. Each eLOA is compared to the iLOA, obtaining the successful and wrong fusions in each case.

³ In the ISSIA soccer dataset it is extracted from the ground truth tracking of each of the 6 cameras. The blob ID for the blobs of a player is the same for the six tracking files.

⁴ The LOA, iLOA and eLOA are “classes”, which are instantiated as specific lists (e.g., FCeLOA, RiLOA) during the different fusions performed (see section 6.2).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

If a blob number is found in the same line in both lists, the fusion is correct. The total number of elements in the iLOA corresponds to the total number of expected fusions. The total number of elements in the eLOA corresponds to the total number of fusions obtained by the system.

After obtaining the necessary data, the values of recall and precision for the fusion process are calculated. Definitions of recall and precision are extracted from [12]: Recall is the fraction of accurate associations to the true number of associations; Precision is the fraction of accurate associations to the total number of achieved associations. Let ξ_{Ω} be the ground truth for pairs of trajectories on the overlapping region and let E_{Ω} be the estimated results. Then, R and P are calculated as:

$$Recall = \frac{|\xi_{\Omega} \cap E_{\Omega}|}{|\xi_{\Omega}|}$$

$$Precision = \frac{|\xi_{\Omega} \cap E_{\Omega}|}{|E_{\Omega}|}$$

6 Adjustments, testing and results⁵

For the testing, the used videos are from ISSIA Soccer Dataset (see section 2.2). The available ground truth tracking consists of an ideal tracking of each camera in which, besides the position and size of each player in each frame, the unique ID of each blob is provided. It allows knowing which player corresponds to each blob. This tracking had some annotation errors that were corrected when they were detected⁶.

There are some important definitions that facilitate understanding the rest of the section:

- Facing cameras: Overlapping cameras covering an area of the field. There are three pairs of facing cameras: camera 1 and camera 2, camera 3 and camera 4, camera 5 and camera 6.

⁵ A web page has been created where some videos with the result of the system have been published. The graphics of Precision and Recall have been added with higher resolution.

<http://www-vpu.eps.uam.es/publications/DetectionAndTrackingInMulticameraSportsVideo/>

⁶ A document with the corrections is available in the created web page.

- **Regions:** The resulting fusion of each pair of facing cameras is named region. There are three regions, one for each pair of facing cameras.
- **Field:** The result of combining the three regions, covering all the field area, by fusing the two overlapping areas of the regions (region of cameras 1-2 and region of cameras 3-4, and region of cameras 3-4 and region of cameras 5-6).

To get the fusion of the field, first facing cameras are fused and then the resulting regions are fused, generating the fusion of the field.

The number of correct fusions for each fusion ground truth is represented in the table 5-1.

Tracking	1 with 2	3 with 4	5 with 6	Total facing cameras fusions	Regions fusion
Ground truth tracking	23	75	39	137	106
Base system tracking	159	462	267	888	132

Table 2: Number of correct fusions

The fusion threshold takes values between 0 and 50. The incremental approach developed for fusion is explained in section 6.1.

There are two different evaluations of the system used, one for each available trackings, which are described in sections 6.3 and 6.4.

6.1 Fusion incremental development

For each of the two available tracking data (the one provided in the ground truth of individual cameras and the resulting tracking of the system with the modifications) the result of the three types of fusion described in section 4 have been analyzed.

The results using basic fusion are presented with the names of GT1 and BS1, depending on the tracking used, Ground Truth (GT) or Base System (BS).

With the basic fusion and thanks to the unique ID of the blobs, to simulate the results to achieve an ideal tracking which discriminate between the different uniforms of the people in the field is possible. The results using this fusion are presented with the names of GT2 and BS2, depending on the tracking used, Ground Truth (GT) or Base System (BS).

Given the results of the first two methods, fusion between cameras 5 and 6 shows worse results than in other pairs. It is due to the homography does not fit properly, for the reasons discussed in section 4.3 as imperfections of the lens or different camera orientation and height. In Figure 9, an example of the problem is shown. Two fragments of the resulting trajectories are presented from the player with unique ID = 104 of the facing cameras 1 and

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

2 (right) and 3 and 4 (left). The vertical distance of trajectories from cameras 5 and 6 is significantly higher than the distance between the trajectories from cameras 1 and 2.

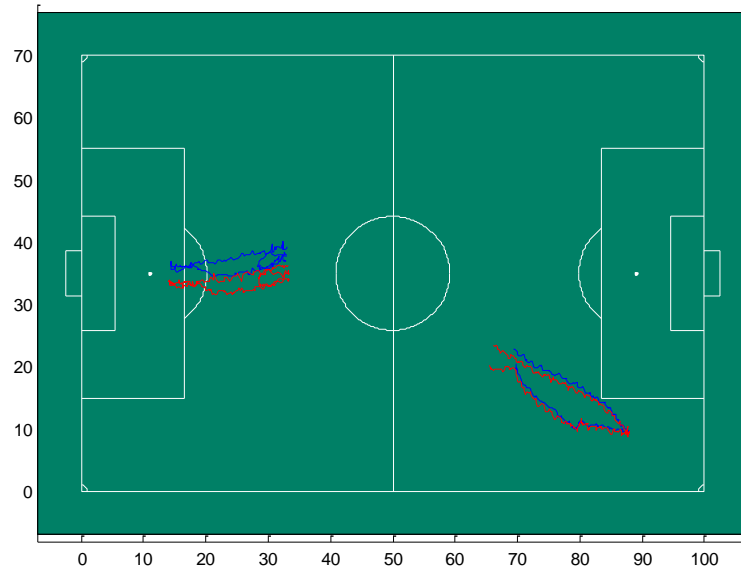


Fig. 9: Example of trajectories before applying the homography without correction

Results named GT3 and BS3 use the same fusion than GT2 and BS2, but correcting the explained error. The example shown in Figure 9 with the correction of the points of the homography is presented in Figure 10.

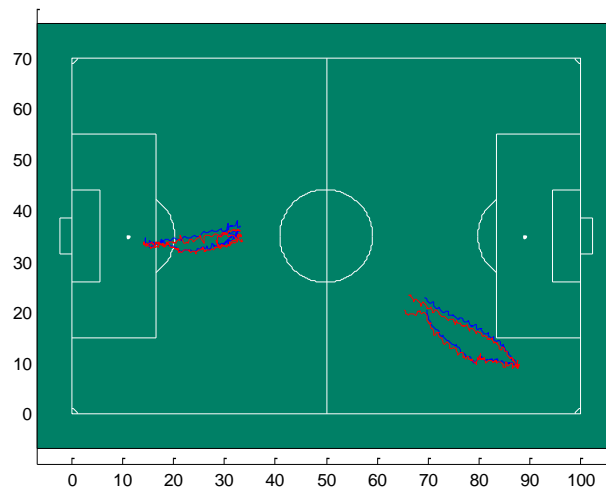


Fig. 10: Example of trajectories after applying the homography with correction

The correction is done heuristically and manually after seeing the results.

In a real system, the correction would be done at initialization. From a warming up video

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

with players moving around the field, the trajectories of players are extracted and the correct fit can be obtained.

The improvement can be seen on the results of the cameras 5 and 6 described below. In the tests, only homographies of the facing cameras 5 and 6 are corrected. As this correction is not compensated in the other facing cameras, the results of regions fusion are slightly worse, but the goal is to show that the correction of the homography can significantly improve the results of the corrected trajectories.

6.2 Steps followed for each validation scenario

The steps for obtaining the fusion of the field, resulting in the trajectories of players, and to get the data needed (instances of LOAs) for the evaluation of the system are presented.

Steps for obtaining the trajectories of players (fusion of the field):

- 1-The annotated blobs are obtained for the different frames after the tracking. In the case of the ground truth tracking, the result is directly the ground truth tracking file, and for the base system tracking the result is the generated output file of the base system.
- 2-The homography corresponding to each of the camera tracking is applied to obtain the top view trajectories of each camera tracking.
- 3-The first experimental lists of blobs associations (eLOAs) are calculated between each two groups of blobs belonging to facing cameras, resulting in Facing Cameras experimental lists of blobs associations (FCeLOAs). There are 3 fusions in total resulting in 3 FCeLOAs, one for each pair of facing cameras.
- 4-The facing cameras blobs are fused according to the obtained fusion lists (FCeLOAs). After this step, the 6 trackings are reduced 3 trackings, one for each region.
- 5-The second fusion lists of blobs associations are calculated between the groups of blobs of the different overlapping regions (region of cameras 1-2 and region of cameras 3-4, and region of cameras 3-4 and region of cameras 5-6. See Figure 13 to see an example with the resulting overlapping areas), resulting in Regions experimental lists of blobs associations (ReLOAs). There are 2 fusions in total resulting in 2 ReLOAs.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

6-The blobs of the different regions are fused according to the calculated ReLOAs. After this step, the final field tracking with all the trajectories is obtained (see Figure 15 to see an example of the resulting field trajectories).

Steps for obtaining the results of the evaluation system:

1-Making use of the FCeLOAs, the evaluation of the facing cameras is made. The ideal lists (FCiLOAs, different for each used tracking due to they depend on the resulting tracking blobs) are calculated with the unique ID (The way to get the unique ID is explained in each scenario: in the ground truth tracking is the true ID contained in the tracking files, and in the base system is obtained with spatial and temporal similarity between blobs of the base system and the ground truth) and compared with FCeLOAs (the lists obtained with the different thresholds and fusions), obtaining Precision and Recall for facing cameras.

2-For the evaluation, in the fourth step for obtaining the regions fusion, the fusion ground truth (ideal list) is used to prevent that the errors in this first stage affect in the evaluation of the next stage. If the ideal fusion is not applied for facing cameras fusion, the second evaluation cannot be calculated. A unique ID cannot be assigned to the resulting blob of the fusion of blobs from different players. Furthermore, when two blobs belonging to a player are not fused in the first stage but are fused in the second stage, two correct fusions are obtained instead of one, which changes the final results. Note that in the process for obtaining the trajectories of the players (without evaluation) the ground truth is not used (ideal lists, unique IDs...).

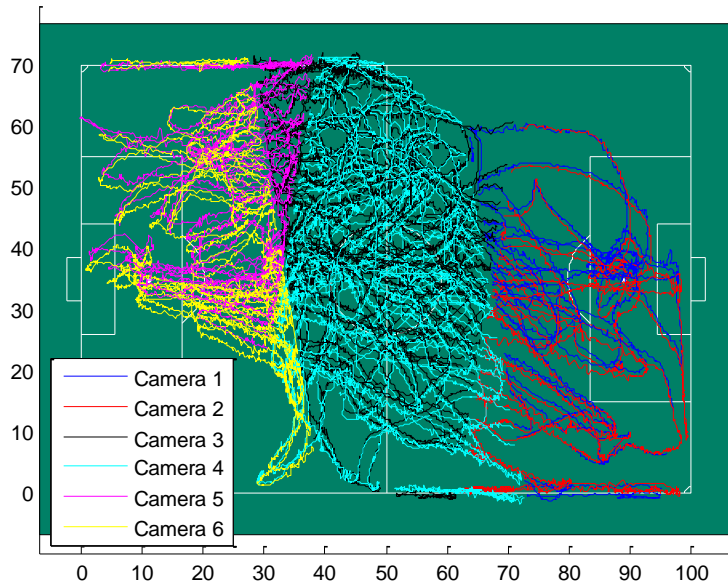
3-Step 1 for obtaining the results of the evaluation system is repeated, but in this case with the RiLOAs and ReLOAs. The result obtained is the precision and recall for the region fusion.

6.3 GT tracking based validation scenario

The first testing uses as tracking the ground truth tracking files provided in the dataset. There are six files, one for each camera, containing the tracking for each player. The blob ID for the blobs of a player is the same for the six tracking files, and is called unique ID. For example, the blobs of the player with unique ID = 5 have that identifier (ID=5) for all the tracking files. The unique blob ID is not used to facilitate the fusions. It is only used to evaluate after the fusion, which allows obtaining Precision and Recall as described in the

1 evaluation system of the section 5. It is used in this way to simulate that the tracking is
2 obtained from a real system and not from the ground truth, with the advantage that quantit-
3 ative results of performance can be obtained.
4

5 The first step is to apply the homography corresponding to each of the six tracking files to
6 obtain the top view trajectories. The result is shown in Figure 11.
7
8
9



32 Fig. 11: Top view tracking for each camera
33

34 Then, the facing cameras fusion is calculated. Results from fusion between facing cameras
35 are shown in Figure 12, which is obtained joining the results obtained from all the facing
36 cameras fusion. The result is not the average from the 3 pairs because the number of fu-
37 sions in each pair of cameras is not the same.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

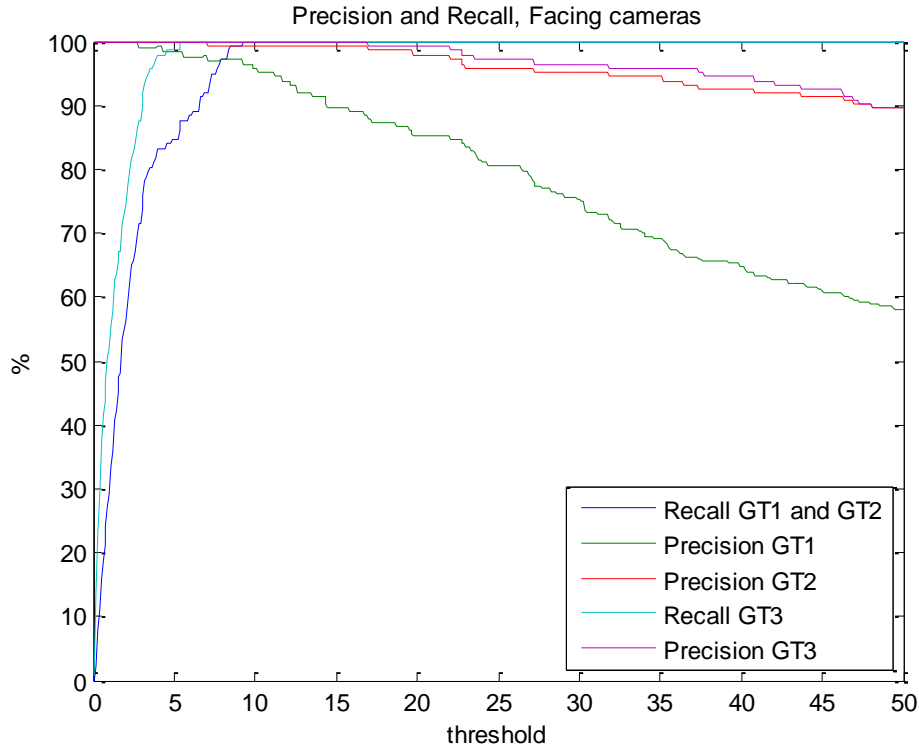


Fig. F-12: Precision and Recall from fusion of facing cameras

The ideal result of the facing cameras fusion is presented in Figure 13. In the figure, each color (red green and blue) represents a region.

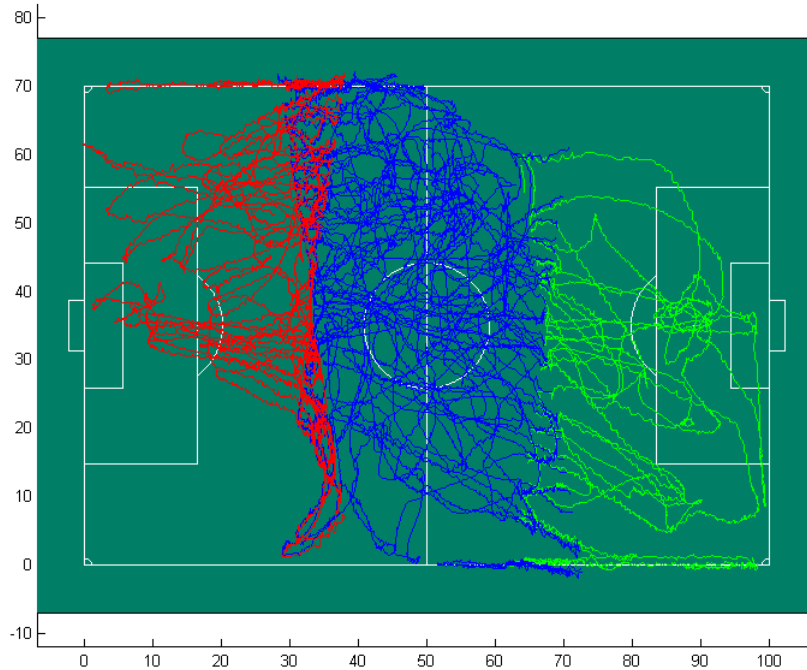


Fig. 13: Resulting trackings from the fusion of facing cameras

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Finally, the regions fusion is calculated. Figure 14 shows the results of fusing the three regions.

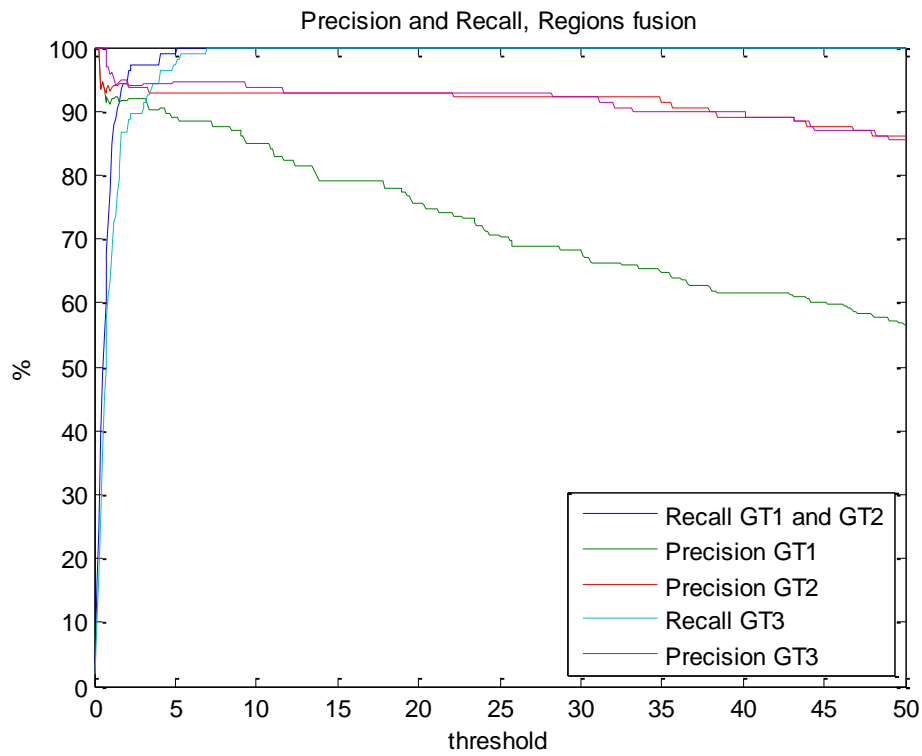


Fig. F-14: Precision and Recall from fusion of the different resulting regions with Ground Truth tracking

The ideal result of the field fusion is presented in Figure 15. In the figure, each player is represented with a different color.

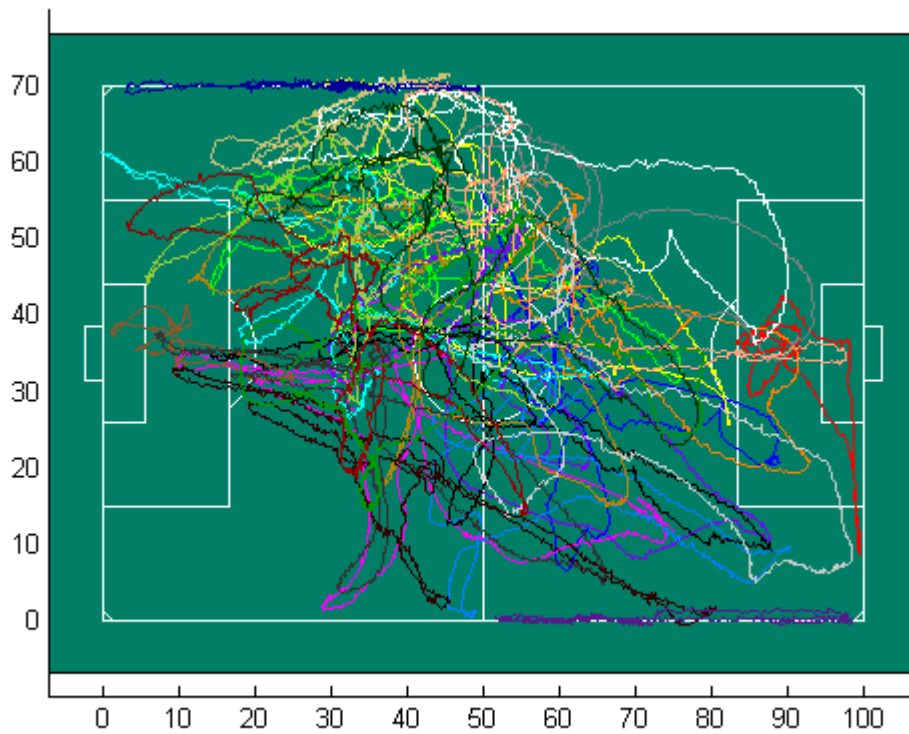


Fig. 15: Resulting tracking from the region fusion

6.4 Base system tracking based validation scenario

The second testing uses the tracking module of the base system⁷. The average time of the tracking processing of the videos (with the applied modifications to the base system) is 7,7 FPS.

The first step is to apply the homography corresponding to each of the six ground truth tracking files and to each of the six base system tracking files to obtain the top view trajectories of both trackings.

To get the unique ID of the base system tracking blobs, the score defined in section 4.1 between the top view ground truth tracking blobs and the top view base system tracking blobs is calculated for each camera. For each blob of the base system tracking module, the corresponding blob of the ground truth tracking with the lowest score indicates the unique ID. Using this method, the identifier of the closest spatially blob (in the corresponding frames) from the ground truth tracking is obtained for each of the blobs obtained from the base system tracking.

After obtaining the unique ID of the blobs of each camera of the base system tracking, the steps 3 to 6 of section 6.2 are applied: FCeLOAs are obtained, the blobs of the facing cameras are fused according to the FCeLOAs, ReLOAs are obtained and finally the blobs of the different regions are fused according to the ReLOAs, resulting in the field trajectories. As in the case of the previous scenario, the unique blob ID is used only to evaluate Precision and Recall after the fusion, as described in the evaluation system of the section 5. Precision and recall for this validation scenario are shown in the following figures. Figure 16 is obtained joining the results obtained from all the facing cameras fusion.

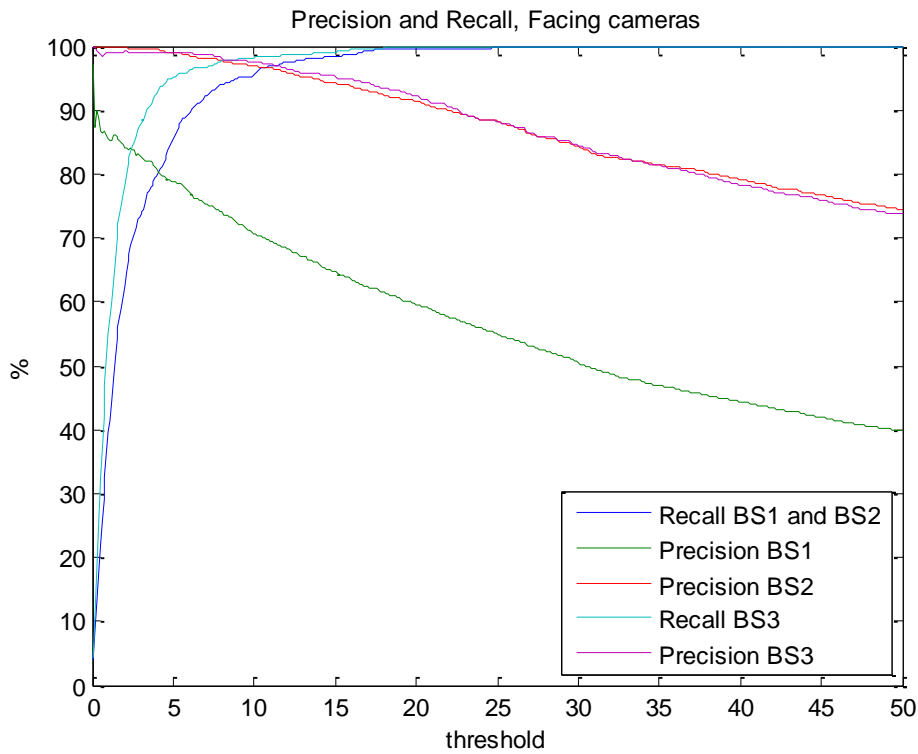


Fig. 16: Precision and Recall from fusion of facing cameras

Figure 17 shows the results of fusing the three regions.

⁷ The base system used is based on the system described in [24].

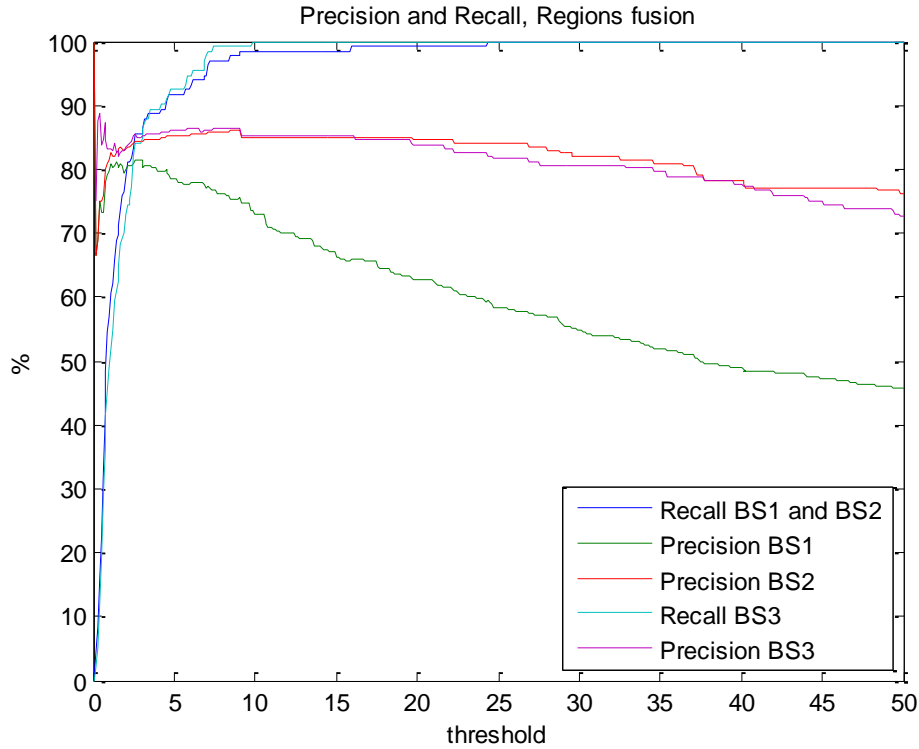


Fig. 17: Precision and Recall from fusion of the different resulting regions

7 Applications

After obtaining the results of the system, some additional functionalities have been implemented to show some players statistics.

A zigzag effect occurs between two consecutive frames, as shown in Figure 18. There are multiple sources of error that cause this problem: camera lenses, homographies, annotation, segmentation, tracking or fusion. This effect causes an error in the obtained statistics because they are higher than the real statistics. To reduce this error, the statistics are calculated every 25 frames, which corresponds to one second of video.

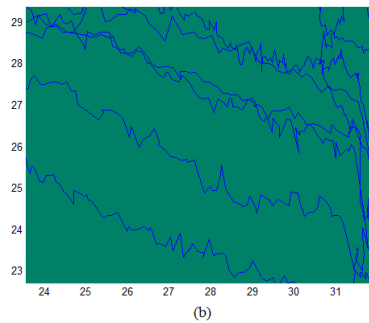


Fig. 18: Example of the zigzag effect in the trajectories

The statistics presented are: position on the X axis, position on the Y axis, covered distance, average speed and instant speed. The position on the X axis and the position on the Y axis are directly the coordinates of the player in a frame. The covered distance is calculated by adding the covered distance in each temporal interval (a temporal interval of one second, as indicated previously). The average speed is calculated by dividing the distance covered by the elapsed time. The instant speed is calculated as the average speed but only considering the last temporal interval.

An example of the resulting statistics video is shown in Figure 19 (the referee tracking is used to facilitate the visual tracking in the video).

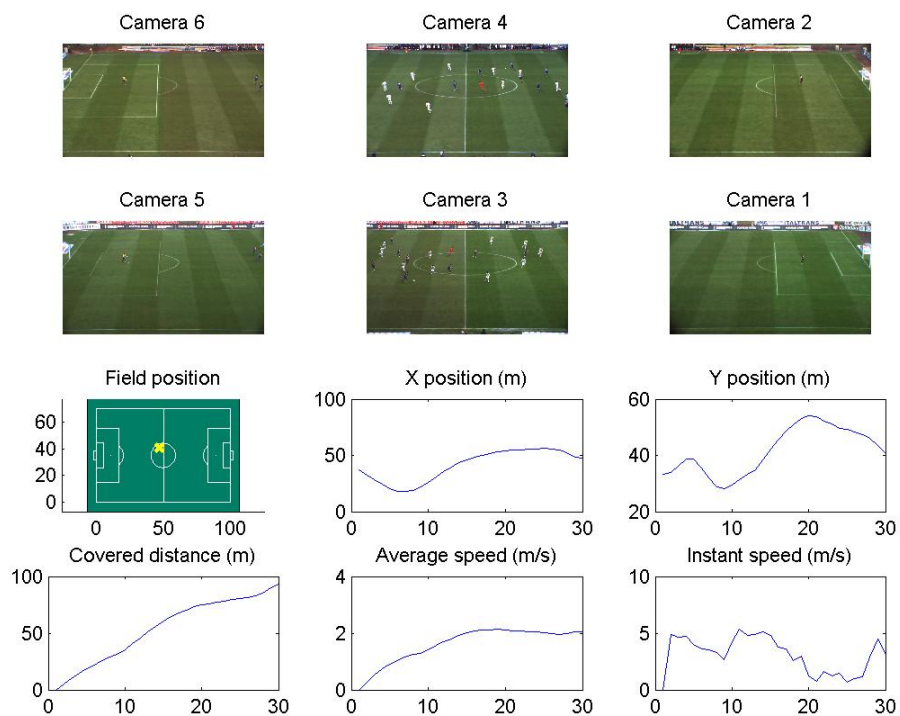


Fig. 19: Example of the resulting statistics video

8 Conclusions and future work

8.1 Conclusions

The main objective of the work presented in this paper, the design and development of a system for detecting and tracking players in a field using multicamera video, has been reached. After a previous configuration and with some supervision, the system is able to

1 detect and track each player in the field, and to provide some statistics. The system is com-
2 plete, general and modular, easing future work improvements and modifications.

3 Sport videos with fixed cameras have significant characteristics that provide advantages
4 for analyzing them with respect to general videos of video surveillance. Backgrounds are
5 generally static and uniform, except for certain areas such as public or dynamic advertis-
6 ing, which can be modeled relatively simple. People tracked have specific and distinct uni-
7 form, at least between the different types of people (players from each team, goalkeepers,
8 referees ...). This last case has a disadvantage, because the players on a team have exactly
9 the same appearance as they wear the same uniform, which may complicate the tracking
10 when occlusions occur.

11 The location of each camera is important. The ideal case is when the cameras are faced or
12 symmetrically placed, since in these cases tracking errors are reduced in fusion process.

13 Better results are shown in cases with greater overlapping. This feature is observed in the
14 case of team sports, when comparing the results of the facing cameras with the results from
15 the fusion of regions.

16 Placing the cameras at higher elevations is interesting because it reduces the tracking error.
17 The greater height of the camera, the smaller area of pixels is projected in the plane of the
18 field and therefore ensures greater precision to the homography projection, but if the cam-
19 era is located too high, the identification of the player uniform can be difficult.

20 In the case of team sports, the main problems are occlusions and regions with low or with-
21 out overlapping. These systems are on which most improvement is needed, as seen in state
22 of art and in experimental results. There are many methods of fusion and many parameters
23 which, when combined properly, may contribute to better results. In the upgrades made,
24 with relatively simple changes, significant improvements are obtained in the results. As
25 shown in the experiments, in the case of the real tracking system a supervised post-
26 processing for obtaining the complete paths is required, whose replacement by an automat-
27 ic process is proposed as future work.

28 **8.2 Future Work**

29 Some future work lines are:

- 30 • *Background extraction*: The extractor used is relatively simple. Work could focus on
31 making an extractor more complex or one with lower computational cost, for example
32 selecting invariant pixels (or within margins) for a certain number of frames.

- 1 • *Graphical User Interface*: As described in the different sections of the system initiali-
2 zation, individual modules are configured manually in the corresponding script. Also
3 for the supervised fusion of partial trajectories (see section 6.4), the tasks is currently
4 done manually. A real-time interactive GUI should be developed for the system opera-
5 tors or system supervisors.
6
7
- 8 • *Tracking system*: This perhaps is the line with more possibilities for development. The
9 system used was slightly adapted from one designed for video surveillance. This sys-
10 tem is a good base but there are many improvements and changes that could produce
11 better results. Some proposals are:
12
 - 13 ○ *Adaptative background*: Taking advantage of the specific characteristics of the
14 background in each sport (public, dynamic advertising, ...)
 - 15 ○ *Detect colors of the uniform of the players*: With this improvement it could be ap-
16 plied the fusion improvement presented and simulated in the team sports fusion
17 section.
 - 18 ○ *Real time*: Designing a system that allows tracking players and fusing trajectories
19 in real time.
- 20 • *Fusion*: As mentioned previously, there are many parameters and methods for the fu-
21 sion of the trajectories in team sports videos. Future work may consist of combining
22 some existing methods or add new ones. The post processing block to automatically
23 connect fragmented resulting trajectories of the players in the team sports system is an-
24 other line of future work.
- 25 • *Performance and Tactic analysis*: From the complete system, additional statistics and
26 information can be calculated. Areas of the field where the team stays longer, place-
27 ment of players on the field, area of influence of each player, etc. Also the effect of
28 zigzag can be studied and corrected.

References

- [1] Y. Xinguo, D. Farin, Current and emerging topics in sports video processing. In Proc. of ICME 2005.
- [2] P. Figueroa, N. Leite, R. Barros, I. Cohen, G. Medioni, Tracking soccer players using the graph representation. In Proc. of ICPR 2004.
- [3] Y. Huang, J. Llach, S. Bhagavathy, Players and ball detection in soccer videos based on color segmentation and shape analysis. In Proc. of ICMCAM 2007.
- [4] C. Poppe, S.D. Bruyne, S. Verstockt, R.V. de Walle, Multi-camera analysis of soccer sequences. In Proc. of AVSS 2010.
- [5] M. Xu, J. Orwell, G. Jones, Tracking football players with multiple cameras. In Proc. of ICIP 2004.
- [6] G. Kayumbi, P.L. Mazzeo, P. Spagnolo, M. Taj, A. Cavallaro, Distributed visual sensing for virtual top-view trajectory generation in football videos. In Proc. of CIVR 2008.
- [7] S. Choi, et al., Where are the ball and players? Soccer game analysis with colorbased tracking and image mosaic. In Proc. of ICAP 1997.
- [8] Tong, X., et al., An Effective and Fast Soccer Ball Detection and Tracking Method. In Proc of ICPR 2004.
- [9] T. Bebie, H. Bieri, SoccerMan: reconstructing soccer games from video sequences, In Proc. of ICPR 1998.
- [10] M. Taj, A. Cavallaro, Multi-camera track-before-detect. In Proc. of ICDSC 2009.
- [11] G. Kayumbi, N. Anjum, A. Cavallaro, Global trajectory reconstruction from distributed visual sensors, In Proc. of ICDSC 2008
- [12] N. Anjum, A. Cavallaro, Trajectory association and fusion across partially overlapping cameras. In Proc. of AVSS 2009.
- [13] Y.A. Sheikh, M. Shah, Trajectory association across multiple airborne cameras, Transactions on Pattern Analysis and Machine Intelligence, 30(2):361-367, Feb. 2008.
- [14] J. Kang, I. Cohen, G. Medioni, Tracking people in crowded scenes across multiple cameras. In Proc. of ACCV 2004.
- [15] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, J.-L. Van Gool, Color-based object tracking in multi-camera environments. In Proc. of DAGM 2003.
- [16] I.N. Junejo, H. Foroosh, Trajectory rectification and path modeling for video surveillance. In Proc. of ICCV 2007.
- [17] I. Sachiko, S. Hideo, Parallel tracking of all soccer players by integrating detected positions in multiple view images. In Proc. of ICPR 2004.
- [18] Y. L. de Meneses, P. Roduit, F. Luisier, J. Jacot, Trajectory analysis for sport and video surveillance, Electronic Letters on Computer Vision and Image Analysis, 5(3):148-156, Mar. 2005.
- [19] R. Hartley, A. Zisserman, A Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.
- [20] P. J. Figueroa, N. J. Leite, R. M. Barros, Tracking soccer players aiming their kinematical motion analysis, Trans. on Computer Vision and Image Understanding, 101(2):122-135, Feb. 2006.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
- [21] T. Misu, S. Gohshi, Y. Izumi, Y. Fujita, M. Naemura, Robust tracking of athletes using multiple features of multiple views. In Proc. of WSCG 2004.
- [22] W. Du, J.B. Hayet, J. Piater, J. Verly, Collaborative multi-camera tracking of athletes in team sports. In Proc. of CVBASE 2006
- [23] T. D’Orazio, M.Leo, N. Mosca, P.Spagnolo, P.L.Mazzeo, A Semi-Automatic System for Ground Truth Generation of Soccer Video Sequences. In Proc. of AVSS 2009
- [24] Juan C. SanMiguel, José M. Martínez, A semantic-based probabilistic approach for real-time video event recognition, Computer Vision and Image Understanding, 116(9): 937-952, Sept. 2012.

PRESUPUESTO

1) Ejecución Material

- Compra de ordenador personal (Software incluido)..... 2.000 €
- Alquiler de impresora láser durante 6 meses 50 €
- Material de oficina 150 €
- Total de ejecución material 2.200 €

2) Gastos generales

- 16 % sobre Ejecución Material 352 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material 132 €

4) Honorarios Proyecto

- 640 horas a 15 € / hora..... 9600 €

5) Material fungible

- Gastos de impresión..... 60 €
- Encuadernación..... 200 €

6) Subtotal del presupuesto

- Subtotal Presupuesto..... 12060 €

7) I.V.A. aplicable

- 16% Subtotal Presupuesto 1929.6 €

8) Total presupuesto

- Total Presupuesto..... 13989,6 €

Madrid, Septiembre de 2012

El Ingeniero Jefe de Proyecto

Fdo.: Rafael Martín Nieto
Ingeniero de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de DETECCIÓN Y SEGUIMIENTO EN VIDEOS DEPORTIVOS MULTICÁMARA. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es

obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.
2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.
3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.
4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.
5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.
6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.