Universitat Politècnica de Catalunya

Master Thesis

Audiovisual framework for automatic soccer highlights generation

Author:
Arnau Raventós

Director:
Dr. Francesc Tarrés

Tutor:
Dr. Ferran Marqués

A thesis submitted in fulfilment of the requirements for the degree of European Master of Science in Research on Information and Communication Technologies (MERIT)

in the

Audio Visual Technologies Group (GTAV)
Department of Signal Theory and Communications (TSC)

June 2015

"Information is the resolution of uncertainty."

Claude Shannon

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Abstract

Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona (ETSETB)
Department of Signal Theory and Communications (TSC)

European Master of Science in Research on Information and Communication Technologies (MERIT)

Audiovisual framework for automatic soccer highlights generation

by Arnau Raventós

Automatic generation of sports highlights from recorded audiovisual content has been object of great interest in recent years. The problem is indeed especially important in the production of second and third division highlights videos where the quantity of raw material is significant and does not contain manual annotations. In this thesis, a new approach for automatic generation of soccer highlights is proposed. The approach is based on the segmentation of the video sequence into shots that will be further analyzed to determine its relevance and interest. For every video shot a set of low and mid level audio-visual descriptors are computed and combined in order to obtain different relevance measures based on empirical knowledge rules. The final summary is generated by selecting those shots with highest interest according to the specifications of the user and the results of relevance measures. The main novelties of this work have been the temporal combination of two shot boundary detectors; the selection of keyframes using motion and color features; the generation of new soccer audio mid-level descriptors; the robust detection of soccer players; the employment of a novel object detection technique to spot goal-posts and finally, the creation of a flexible and user-friendly highlight generation framework. The thesis is mainly devoted to the description of the global visual segmentation module, the selection of audiovisual descriptors and the general scheme for evaluating the measures of relevance. Several results have been produced using real soccer video sequences that prove the validity of the proposed framework.

Acknowledgements

I would like to express the deepest appreciation to...

Francesc Tarrés; for all the wisdom, enthusiasm, motivation and opportunities he has brought to me.

Lluís Torres; for the useful insights and advices that allowed me to see things differently and improve.

Raúl Quijada; my brother in arms, for all the help and good moments that we have shared doing the project.

Javier Rubio; because if it weren't for him I wouldn't be here.

And last but not least, to Mercè and my parents Pere and Maite, for all their love and support.

Contents

List of Figures

List of Tables

Introduction

The production of multimedia content has increased tremendously in recent years. Millions of images and video sequences are created every day and it can be stated that the total number of stored data in any digital support (computers, memory cards, the cloud, etc.) is hardly estimable[1]. Therefore, there is an increasing need to provide advanced tools to automate and facilitate its management [2][3][4][5].

In this myriad of multimedia material, sport video sequences play an important role, as a large variety of sports are played and recorded all over the world getting the interest of many people. However, sports video sequences tend to be lengthy and once the main live event ends people are usually only interested in the most significant parts of it. That is one of the main reasons why generation of automatic highlights of video has been a major area of research over the last two decades [2][6][7][8][9].

Generation of automatic highlights of video sequences consists in creating a shorter video version that retains the most interesting parts of the original sequence while disregarding those events that may be considered of low interest by the viewer. In the context of sport broadcasting or streaming, automatic systems that produce highlights of a match may play a key role in the overall cost of the production chain.

Many TV stations need to have summaries of video sport sequences in order to broadcast them once the matches have been finished. Usually, the sequences are analyzed and summarized manually by a journalist and this operation represents a laborious and exhausting task. Due to the great amount of available sequences and the amount of time required in this process, there is a necessity to provide tools that speed this summarization process. In addition, being soccer extremely popular in many countries, it is one of the most important areas where sports video sequences summarization is being applied. A variety of approaches have been presented in the literature.

[1]Stored data estimation: *http://www.insideactivitytracking.com/data-science-activity-tracking-and-the-battle-for-the-worlds-top-sports-brand/*

For many years, low-level descriptors have been the only approach available for soccer video sequences summarization and in general for video analysis. These low-level descriptors include, among others, statistical moments, shape, color, texture, and motion. However, it is well recognized that such information is not enough for uniquely discriminating across different visual content. The use of advanced information is required in order to obtain meaningful results. Several different approaches have been introduced in video summarization. Among them, spatio-temporal analysis, video structure and syntax and video sequence events have gained a lot of attention [10][11][12][13][14][15].

A short review on the state of the art of soccer video summarization is introduced in the next paragraphs.

Ekin et al. present in [16] a fully automatic and computationally efficient framework that classifies shots taking into account *cinematic* and *object-based* features. Cinematic features refer to those that result from common video composition and production rules, such as shot types and replays and they are highly efficient to compute. On the contrary, object-based features are spatial (color, texture, and shape) and spatio-temporal features of objects and usually have a higher computational cost. The approach followed by the authors is to only compute object-based features when the cinematic features are not enough to detect an event. With these two type features the system is able to classify shots and detect slow-motion replays, goals, the referee and the penalty box. The framework provides three types of summaries: All slow-motion segments, all goal events, and extensions of the first two.

O'Connor et al. define in [17] an event detection framework that can be applied in a variety of sports genres. These sports share a set of common characteristics such as two opposing teams plus referees, an enclosed playing area, a grass pitch, field lines, a commentator voice-over, a spectator cheering...Their goal is to model which events occur in all these sport-type situations and they find that important events are followed in 98% of the cases by an action replay. Thus, the system aims to detect these time instants by introducing the concept of a *reaction-phase*, which is the lag time that immediately follows the occurrence of an event, before the cut to action replay. A reaction-phase content typically includes a close-up, a celebration, an audio increase, an on-screen graphics or a near-field motion activity. Consequently, a definition on how to extract features from these situations is presented and finally a support vector machine is trained to detect them.

Zawbaa et al. present in [18] a machine learning based event detection and summarization system. It firstly segments the whole video sequence into video shots, then it classifies the resulted shots into different shot-type classes. Afterwards, the system applies a support vector machine and an artificial neural network algorithm for identifying

important segments with replays. Independently, the system detects vertical goal posts and the goal net. Finally, all the information is combined and the most important events during the match are highlighted following a set of rules.

Tavassolipour et al. introduce in [19] a soccer summarization framework that is able to detect goals, cards, goal attempts, corners, fouls, offsides and no-highlights. It initially segments the video in shots and then it automatically detects replays. Later on shots are classified in several views and a hidden markov model is used to group them and create larger semantical units termed *play-breaks*. These units consist of two sections called *play* and *break*. In soccer, the game is in a *play* mode when the game is going on and the *break* mode is the complement set; that is, whenever the game is halted because of occurrence of an event. A set of low-level features are extracted from these units and afterwards a bayesian network is built to classify the soccer events. Finally, the summarization process is treated as a knapsack problem and solved using dynamic programming.

It is in this context that a new approach for soccer video sequences highlights summarization is presented. The approach presented here is based on sport video edition human-expertise used in commercial television. The aim of this work is to automatically detect soccer highlights and at the same time provide an adjustable and accessible framework for generating the soccer summary. The system defines the shot as the minimum unit for building up the summary. A video segmentation approach is introduced that relies on the detection of shot transitions. Then, a score is assigned to every shot and those shots with the highest score are selected in order to build up the final summary. To qualify for scoring, shots are first passed through an analysis bank that computes several low-level and mid-level audiovisual descriptors. These descriptors define the video contents for every shot. Once the shot descriptors have been computed, they are passed through a multiple filtering process that will attempt to associate semantic meaning to each shot and will provide an overall score. The scheme is designed such as the user is able to specify, up to some degree, the type of contents appearing in the summary and its approximate duration.

The main novelties of this work have been the temporal combination of two shot boundary detectors depending on the time instant of the match; the selection of keyframes using motion and color features; the generation of new soccer audio mid-level descriptors; the robust detection of soccer players; the employment of a novel object detection technique to spot goal-posts and finally, the creation of a flexible and user-friendly highlight generation framework.

This work has been part of the BUSCAMEDIA research project [1] founded by the Spanish National Science Foundation made in cooperation with TVC, the Catalan TV

broadcaster. Several members from the Audio Visual Technologies Group (GTAV) participated in the project and each one of them focused in one or more tasks. In this thesis I will describe the entire system to provide the reader a global view of the framework, even though I will empathize the parts where I actively contributed. Specifically these parts are: The visual segmentation module (Shot Boundary Detection, KeyFrame Selection), a set of description tools (Long Shot Detector, Zoom Detector, Goal Post Detector) and the Highlight Generation Framework. The remaining description tools (Person Detector, Replay Detector, Whistle Detector and Inter and Intra shod-based audio power detectors) were developed by others and will be briefly depicted in Section 3.5.

This thesis is divided into the following chapters. Chapter 1 presents an overview of the overall system. This section depicts the philosophy behind the work-flow of all parts in the automatic soccer highlight generation framework. The segmentation of the video sequence in minor units is depicted in Chapter 2 with the description of the shot boundary detectors and the keyframe selector. Later on the details of the low-level and mid-level audiovisual descriptors is specified in Chapter 3. Chapter 4 introduces the approach for filtering shots for event interest selection in soccer. The chapter describes how descriptors can be combined to detect events and how an end-user can easily adjust the percentage of appearance of each one of these events in order to generate the final summary. In addition the results and performance of the system are assessed by analyzing two soccer summaries. Finally, Chapter 5 contains the conclusions and future work.

Chapter 1

Overview of the system

In our approach the design of the automatic soccer game highlights summary generator is based on sport video edition human-expertise used in commercial television. TV sport summaries are made-up of a sequence of short shots that collect the essential events of interest, usually in a linear timely basis where different shots are presented in the same order that have occurred. It is also assumed that the summary is generated from the video feed that was produced during the live broadcasting of the game. No auxiliary cameras, player-follow shots or alternative views are supposed to be available to generate the highlights video summary.

Keeping this TV production style in mind, the minimum unit for building up the highlights video summary will be the video-shot, defined as a series of continuous frames captured by a single camera that runs for a period of time. The final summary will be a concatenation of selected video-shots found in the original sequence. The overall diagram of the automatic highlights generator is represented in Figure 1.1. The objective is to assign a score to every shot and then select those shots with the highest score in order to build up the final summary. To qualify for scoring, shots are passed through an analysis bank that computes several low-level and mid-level audiovisual descriptors. Once the shot descriptors have been computed, the resulting xml files are finally passed to the highlights generation stage which, through a multiple filtering process, will attempt to associate semantic meaning to each shot and will provide an overall score. The final score of a shot may take into account not only its own descriptors but also the ones of its neighbor's shots. The user may interact with the system specifying the total duration and the percentage of every semantic filter in the final summary.

One of the key components in the architecture of the system is the video-shot segmentation module that produces an output xml file indicating the initial and ending time codes for every shot. The output of the video-shot segmentation stage is then used by

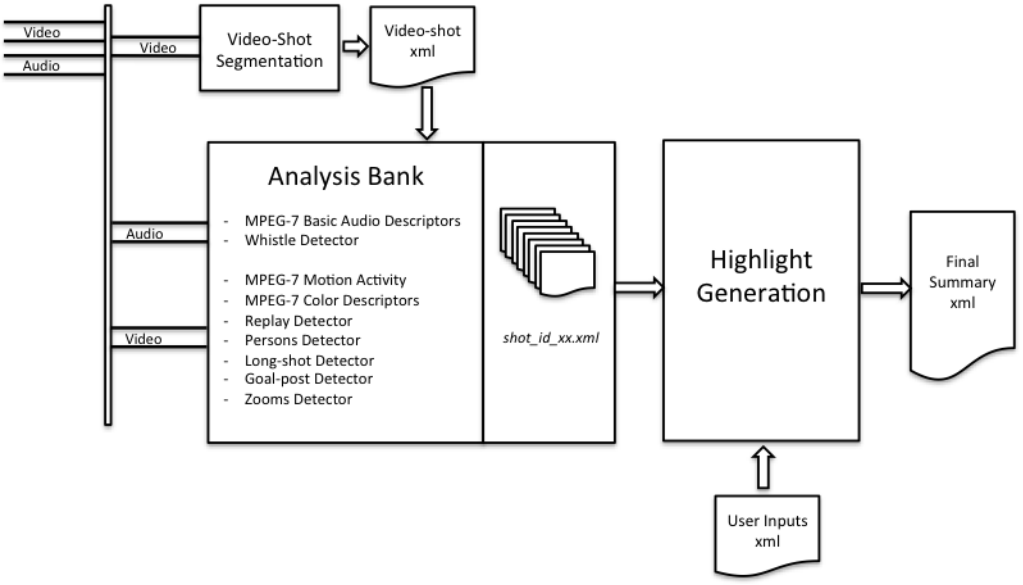

FIGURE 1.1: General diagram of the automatic soccer highlight generator

the analysis bank to process the audio and video tracks of every shot and determines a set of audio and video descriptors that will be annotated in xml format and transferred to the highlights generation module which will generate the final highlights video summary. In addition, another key component of the summarization system is the use of the audio information that, combined with the video information, provides more robust and accurate results. In particular, a new algorithm to extract the whistle information found in the audio track is presented in the Subsection 3.5.3 and the employed MPEG-7 basic descriptors are presented in Subsection 3.1.

The reliability of the video-shot segmentation is based on the detection of the transitions. In live sport broadcasting, transitions are mainly hard-cuts and cross dissolves. Hard-cuts are used during the action of the game while the latter may be used before or after the game or during the half time. The algorithm for the transition detection is explained in Subsection 2.1.2 and combines two off-the-shelf methods in order to achieve a good trade-off between complexity and performance with these two types of transitions, selecting one or other method depending on the part of the game video sequence that is being processed.

The analysis bank is represented in Figure 1.2 and includes a variety of analysis tools to extract the audio and video descriptors. Each of these tools produce an elementary xml file that describes the contents of the audio or and video sources with a single descriptor type. Then, these elementary xml files are parsed together to obtain a set of xml files collecting all the classes of descriptors for every shot in the video sequence. Some of

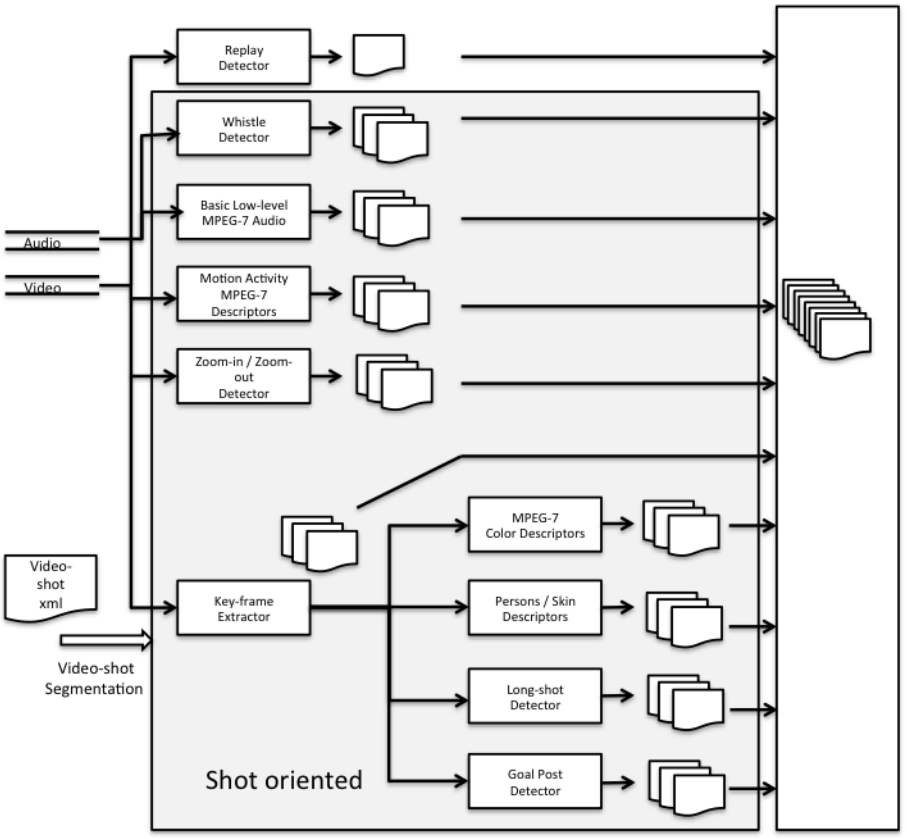

FIGURE 1.2: Schematic representation of the analysis bank

these tools are applied directly to the segmented video or audio streams while others will only process a set of selected key-frames of the video sequence. In the latter case, the analysis tools are used at the output of a key-frame detector that selects a few frames (in most cases a single frame) that represent the shot.

The information collected from every shot includes:

- Low-level audio MPEG-7 descriptors for every audio frame.
- The motion descriptors (Motion Activity and Camera Motion) associated to the shot.
- Detection of Zoom-in and Zoom-outs in the analyzed video shot.
- A list of key-frames that represent the video shot.
- Color Layout, Color Structure and Dominant Color MPEG-7 descriptors associated to every key-frame in the shot.
- Faces detected in the key-frames. Percentage of skin pixels in a key-frame.

- Long-shots detected in the key-frames.
- Goal-posts detected in the key-frames.
- Replays detected in the broadcasted audio feed time codes.
- Referee's whistle time codes.

The details of the algorithms used in each module are explained in Chapter 3.

The final stage of the system is the highlights summary generator module, that is composed of a filter bank that scores each shot with a set of elementary and advanced filters and the shot selection stage that generates the final summary by assessing the previous scores and a set of user parameters. The elementary filters find important characteristics of the low-level descriptors in the video shot. For example, one elementary filter can be the audio power increase inside the shot or the detection of the referee's whistle. Then, elementary filters are combined together in order to create a set of advanced filters that will give a final score to every shot. The objective is to assign positive or negative weights to each elementary filter in order to obtain the final score assigned to every shot. These filters are flexible enough to let the user interact with the system, specifying the relative importance of each elementary filter or, alternatively, to be statistically trained by a linear classifier. The user may define the number of advanced filters, their structure, the weighting of elementary filters, the percentage of every advanced filter included in the summary and the summary duration. The highlights summary generator module is presented in detail in Chapter 4.

In order to evaluate the performance of the proposed approaches and the overall automatic highlights summary generator, three different video groundtruths have been generated. Each groundtruth contains the following soccer matches:

- Groundtruth A: Three matches from 1st division and two matches from 2nd division.
- Groundtruth B: Three matches from 2nd division and three matches from 3rd division.
- Groundtruth C: One match from 1st division, one match from 2nd division and one match from 3rd division.

The idea behind splitting the groundtruths taking into a account the division of the soccer matches is due to their differences in production styles and luminance conditions. As it is very difficult to generate a feature detector that works well in all scenarios, there

would be descriptors specially designed for concrete divisions and thus their performance will only be tested in that specific environment. For instance, the proposed long-shot detector pretends to recognize those shots where a predominant number of pixels share a grass-like color. This approach is suitable for first and second divisions, however, it is not applicable in third division because in the majority of the matches the pitch has large gaps with no grass and the luminance conditions are extremely bad.

Groundtruth A is proposed for those descriptors where the lighting conditions are good and the match is broadcasted using a great number of cameras. Descriptors that use this database are detectors of zooms, goal-posts, persons and long-shots. On the other hand, groundtruth B is aimed for those divisions that sometimes lack many cameras but other features can be exploited. The replay detector based on logo-matching employs this database. Finally, groundtruth C is focused on those techniques that must be applied to all types of video sequences, independently of the content. This groundtruth assesses shot boundary detectors and keyframe selectors.

All the video sequences have been produced by the Catalan television (TVC) and they belong to the Spanish soccer leagues. Each match lasts approximately two hours and is in MPEG-4 Simple Profile format, has a resolution of 360x288 and a GOP structure IPPPPPPP. All the results presented in the sequel have been obtained using these video groundtruths.

Chapter 2

Visual Segmentation

In order to generate a video summary, the original video sequence is usually split into shorter visual temporal units. This allows a structured analysis enabling an easier and faster understanding of what's happening inside the video sequence. Depending on the type of summary these temporal units can range from a video scene up to a single frame [2][8][11]. A temporal video segmentation schematic is depicted in Figure 2.1. In this framework, the proposed basic unit is the video shot and each one of them will be represented by its timecodes and a set of keyframes. The end goal of this methodology is to be able to annotate the whole video sequence by describing the video shots using low and mid level audio-visual descriptors.

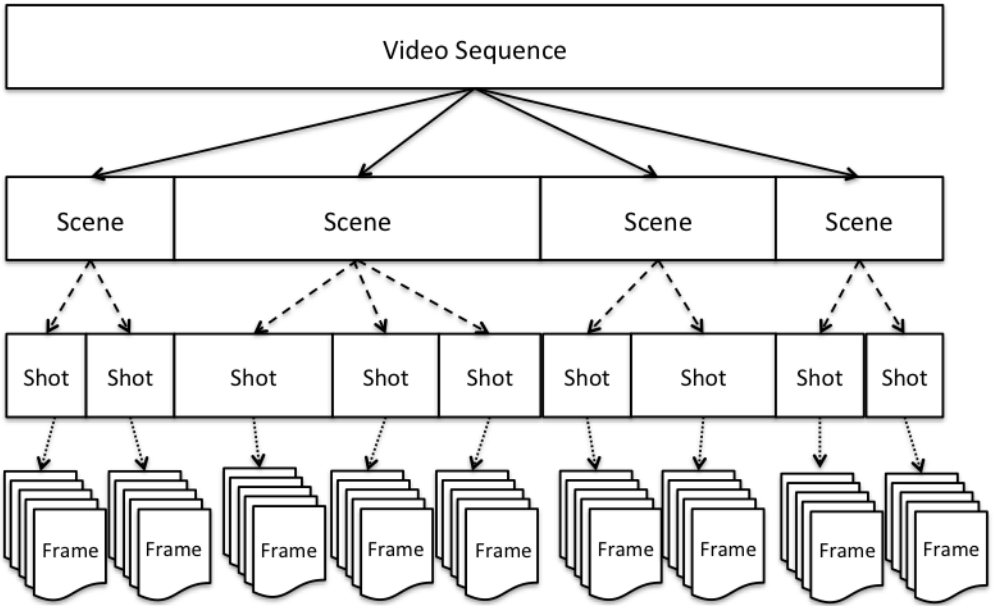

FIGURE 2.1: Temporal video segmentation schematic

In this chapter several approaches for shot boundary detection are implemented and evaluated using soccer matches. Afterwards an special combination of them is suggested for the proposed automatic highlight generator framework. Similarly various methods for keyframe selection are tested and a novel approach is introduced.

2.1 Shot Boundary Detection

A shot is defined as a series of consecutive frames taken contiguously by a single camera and representing a continuous action in time and space. Transitions between shots can be categorized into to two main types: Abrupt transitions and gradual transitions. In abrupt transitions the shot changes completely in one time instant and in gradual transitions the change lasts more than one frame due to a concrete editing effect, such as a dissolve, wipe, fade out/in, etc. The goal of shot boundary detection algorithms is to identify shot transitions in a video sequence. To do so the idea is to extract representative features from the video frames and find a discontinuity in those features along the time domain. The main problem [21] is that there's the necessity to find features that share two qualities: invariance and sensitivity. The invariance because the feature needs to be stable to some forms of content variation except shot transitions, e.g., rotation or translation of the picture. But at the same time they must have a high sensitivity to capture enough details of the visual content so not a single shot boundary is missed. To deal with this problem a lot of methods to perform shot boundary detection have appeared in the literature [2][20][21][22] and even specific approaches have been proposed for soccer games applications [15][17][18][19].

An important number of methods [22] for shot boundary detection try to cope with the problem of finding the best features that offer the greatest trade-off between invariance and sensitivity in the spatial and the time domain. Their objective is to be able to detect both gradual and abrupt transitions. Though the abrupt transition problem is mainly resolved, the detection of gradual transitions may be an issue in some cases.

Some of the most known classical approaches are the following: Direct pixel difference between adjacent frames, which is really fast computationally but offers strong sensitivity in the spatial domain that origins problems in high motion scenes. Color histograms comparison, one of the most frequently used for its good trade-off between the sensitivity in the spatial domain and invariance in the time domain. This approach yields significant results even though it has some difficulties with shots that share similar color PDFs. Another method is to use motion compensation vectors, which usually requires high computational time for their correct extraction although less precise approaches extract them directly from the MPEG streams. Motion vectors assessment shows also

complications with high motion shots. Finally, image edge detections are also commonly employed for shot boundary identification, but require high computational costs and do not demonstrate great improvement with respect to the other methods.

The main difficulties encountered in shot boundary detection applied to soccer live broadcasting events are the high color resemblance of soccer shots, the randomness of their motion statistics and the mixture of abrupt and gradual transitions. Bearing that in mind, the approach presented here consists in employing two types of shot boundary detectors, one for hard transitions and the other for gradual transitions. In soccer live broadcasting, hard transitions are used during the action of the game while cross-dissolves may be used before or after the game or during the half time. Consequently, the proposed approach consists in detecting the game boundaries with the help of a whistle detector algorithm that determines the beginning and ending of each half time. The whistle detector is presented in the Subsection 3.5.3. Once the initial and final whistles are spotted in the temporal domain the abrupt shot detector is employed during the action of the game and the gradual transition detector in the beginning, half-time and final parts of the game.

In the following subsections two proposed shot boundary detectors for hard transitions are assessed and a gradual transition detector is presented. The precision and recall metrics used to test the performance of the detectors are detailed in Appendix A.

2.1.1 Hard transitions

In this subsection two hard transition detectors are evaluated. First a classical state-of-the-art method based on comparing color histograms is introduced and then a novel approach suggested by M.Omidyeganeh et al.[51] based on parametrizing the PDF of wavelet domain coefficients is presented. This second approach is tested because it is a novel strategy that claims to surpass all existing shot-boundary detectors in the well-known TRECVID[1] 2006 shot boundary detection task dataset. Both detectors are compared with a common groundtruth, the results are analyzed and one of them is chosen for the proposed framework.

2.1.1.1 Color histogram based

The first hard transition detector for the suggested framework consists on the traditional comparison of the frame-by-frame distance using as main feature the frame's color histograms. The main goal of the detector is to find temporal dissimilarities in the frames'

[1]TREC Video Retrieval Evaluation website: *http://www-nlpir.nist.gov/projects/trecvid*

PDFs. To do so, frames are converted to the Hue Saturation Value (HSV) color model and then unidimensional histograms are computed for each component. 48 bins are used as features, 32 belonging to Hue, 8 to Saturation and 8 to Value. The HSV color model is employed for measuring the distance between frames because it offers more similar results to the human visual system than the RGB color model. In addition, more bins are dedicated to the Hue component because is the channel focused on color tone and it is desired to have more sensitivity there. With the features extracted from the frames, the distance used to compare them is the Chi-square (2.1), which is defined as follows:

$$d(\mathbf{h}_1, \mathbf{h}_2) = \sum_{i=1}^{N} \frac{(h_1(i) - h_2(i))^2}{h_1(i)} \tag{2.1}$$

where the N stands for the total number of histogram bins and $\mathbf{h_1}$ and $\mathbf{h_2}$ are the histogram features computed from two consecutive frames.

The distances between all adjacent frames are stored in a vector. This vector is normalized by subtracting its mean and then dividing it by its variance. In the end, shot boundaries are detected by comparing the normalized distances with a global predefined threshold. This approach yields an efficient computation while achieving very good results. In particular, for the groundtruth C, applying the method only to the action parts of the soccer game, the parts where the match is on, the results depicted in Table 2.1 are obtained.

	Recall	Precision
Color Histograms	95.2%	98.8%

TABLE 2.1: Shot Boundary Detector Results for Hard-Cut Transitions using Color Histograms

2.1.1.2 Parametrized PDFs of wavelet domain coefficients based

The second approach for hard transition detection is a novel method [51] presented by M.Omidyeganeh et al. The objective of their work is to extract frame-by-frame distances by parametrizing the PDFs of wavelet domain coefficients. First the details of each frame are captured in a multi-resolution manner with the wavelet transform. Then the PDFs of the wavelet high-pass subbands are parametrized with Generalized Gaussian Density functions (GGD) (2.2). Finally the distances among adjacent frames are computed with the Kullback-Leibler distance (KLD) (2.3).

$$p(x, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} e^{-(\frac{|x|}{\alpha})^{\beta}} \tag{2.2}$$

where Γ is the Gamma function, α models the peak of the GGD and is called the scale parameter; and β is proportional to the inverse of the decreasing rate of the PDF and is the shape parameter.

$$D(\mathbf{p}||\mathbf{q}) = \sum_i p(i) ln \frac{p(i)}{q(i)} \tag{2.3}$$

where $\mathbf{p}$ and $\mathbf{q}$ are two probability density functions.

Wavelets coefficients have too much sensitivity in the spatial domain and that's one reason for using as features the PDFs of the subband coefficients, to sacrifice some sensitivity and obtain some invariance in time. Moreover, the parametrization of the PDF of the wavelet transform of an image into a GGD function was introduced in [52] and experimental results showed that the GGD function could yield a reasonable approximation to the marginal statistics of 2D wavelet coefficients (see Figure 2.2). Apart from that, another advantage of parametrizing the PDFs is the reduction in the problem's dimensionality. Instead of storing and comparing all the histograms per subband, only two parameters that approximate all this information are employed. Once

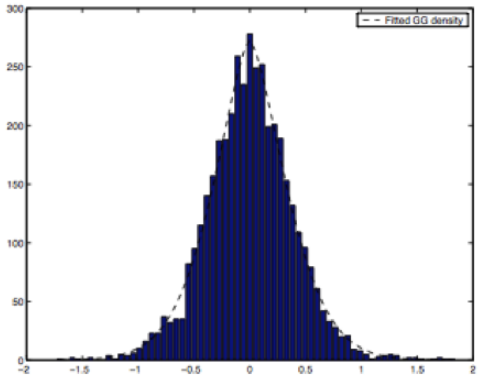

FIGURE 2.2: PDF fitted in a GGD function

the PDFs are fitted in a GGD function, the distance among them is computationally inexpensive, because being the GGD function a closed expression, if (2.2) and (2.3) are combined then a specific KLD (2.4) equation is given by:

$$D(p(., \alpha_i, \beta_i)||p(., \alpha_j, \beta_j)) = log(\frac{\beta_i \alpha_j \Gamma(\frac{1}{\beta_j})}{\beta_j \alpha_i \Gamma(\frac{1}{\beta_i})}) + (\frac{\alpha_i}{\alpha_j})^{\beta_i} \frac{\Gamma(\frac{\beta_j+1}{\beta_i})}{\Gamma(\frac{1}{\beta_i})} - \frac{1}{\beta_i} \tag{2.4}$$

In addition, considering the realistic assumption that coefficients in different subbands of a concrete level are independent, the total distance between two adjacent frames can be computed as the sum of the KLDs between all the equivalent subbands:

$$D(\mathbf{fv_i}, \mathbf{fv_j}) = \sum_{l=1}^{3L} D(p(., \alpha_i^l, \beta_i^l)||p(., \alpha_j^l, \beta_j^l)) \tag{2.5}$$

where L is the number of wavelet transform levels, **fv** is a feature vector containing all the α and β parameters per frame, and i and j are the indices of the frames.

In this manner (2.5) a KLD vector containing all the distances between adjacent frames is built. At this point shot boundaries are detected applying an adaptive thresholding technique. The threshold is defined as $T_s = pm_w$ where p is an empirical number in the range of 2-3 and m_w is the local mean of the KLD vector with a length between 4-10. An example of this adaptive thresholding technique can be seen in Figure 2.3.

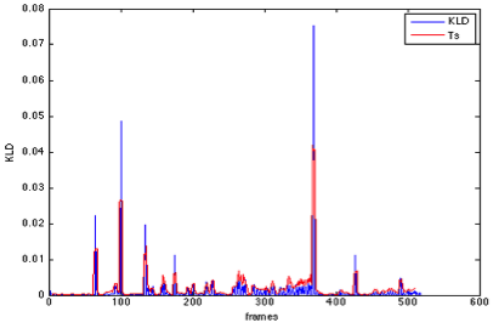

FIGURE 2.3: Adaptive thresholding in a KLD vector.

To implement the proposed approach several wavelets are tested with resolutions of 5 levels, two p values are assessed and a local mean m_w of length 10 is employed. The tested wavelets are asymmetric and orthogonal wavelets like the Haar, Dauchebies-5 and Symlets-5; symmetric and orthogonal like the discrete Meyer and symmetric and biorthogonal like the Biorthogonal-5.5. The same modified groundtruch C of the previous detector is used. Results are represented in Table 2.2.

2.1.1.3 Analyzing the hard transition detectors' results

Remarkably the state-of-art frame-by-frame color histogram comparison provides exceptional results with a quite simple algorithm. Being the missing shots either gradual transitions or very similar shots in terms of color. On the other hand, the results presented by M.Omidyeganeh et al. in [51] could not be reproduced. Gradual transitions

	p=2.7		p=3	
Wavelet	Recall	Precision	Recall	Precision
Discrete Mayer	80.5%	85.2%	75.8%	97.8%
Symlets 5	77.8%	84.3%	74.9%	95.6%
Daubechies 5	79.7%	84.2%	75.7%	92.3%
Haar	80.7%	84.9%	77.8%	93.4%
Biorthogonal 5.5	80.3%	84.4%	76.8%	95.9%

TABLE 2.2: Shot Boundary Detector Results for Hard-Cut Transitions using Parametrized PDFs of wavelet domain coefficients

are also not detected and adjacent shots predominated with grass-type texture are not discriminated.

It can be noted that in our implementation of [51] different wavelets provide almost the same outcomes. In our opinion this could mean that although each wavelet characterizes differently the texture of a frame and a specific type of wavelet could offer better scores for a concrete texture, in general, the matching between distinctive textures with different wavelets provides similar results. In fact this confirms what Brunner and Kadiyala [6] stated with their study of over twenty types of wavelet basis for texture classification. They exposed that the maximum difference in classification rate from the best wavelet basis to the worst was less than 3%.

Apart from the results, computationally speaking the first detector is much faster than the second. Calculating the wavelet transform in 5 levels and parametrizing each subband requires a significant higher cost than transforming the color model from RGB to HSV and computing its histogram.

To sum up, the second approach was discarded and after evaluating the results and the computational costs of the color histogram comparison, we decided that this was a very adequate approach for detecting hard transitions in the proposed framework.

2.1.2 Gradual transitions

Usually in the beginning, half-time and ending parts of a soccer game less relevant events occur in front of the cameras. Typically the reporter of the match speaks in voice-over while the director is showing how the players are entering/leaving the pitch and the audience is performing the initial/final cheerings. It is in these time periods that usually gradual transitions appear in the video sequence. Thus, knowing that the previous algorithms are not designed to detect this type of transitions due to the low variation of the features among adjacent frames, a different shot boundary detector is required.

The main approach followed by gradual transition detectors is to analyze the distance in-between a continuous set of frames instead of only focusing on frame-by-frame changes. However, a good gradual transition detector needs to discriminate among a high-motion camera displacement and a gradual shot boundary transition, and that's a complex non-solved task.

The chosen cross-dissolve boundary detector has been proposed by W. Abd-Almageed in [23] and is based on the Singular Value Decomposition (SVD) analysis of a frame sequence. The technique consists in evaluating the rank of a feature matrix $\mathbf{X}^t$ at the time index t. This feature matrix is expressed as a sliding-window of N frames filled with the Hue Saturation Value (HSV) histograms. To define it, let us denote one frame of this window as a time-varying feature vector $\mathbf{x}^t$ given by this (2.6) formula:

$$\mathbf{x}^t = [\mathbf{h}_H \ \mathbf{h}_S \ \mathbf{h}_V] \tag{2.6}$$

where $\mathbf{h}_H$, $\mathbf{h}_S$, $\mathbf{h}_V$ are row vector histograms and the length of the vector $\mathbf{x}^t$ is L, the sum of the lengths of the histograms.

And let us assume that a set of N continuous frames from each sliding-window fill a NxL feature matrix $\mathbf{X}^t$ in the following manner:

$$\mathbf{X}^t = \begin{bmatrix} \mathbf{x}^t \\ \mathbf{x}^{t-1} \\ \vdots \\ \mathbf{x}^{t-N+1} \end{bmatrix} \tag{2.7}$$

where $t = N, ...T$ and T is the total number of video frames.

The SVD is applied to factorize the matrix $\mathbf{X}^t$ as shown in equation (2.8).

$$\mathbf{X}^t = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^t \tag{2.8}$$

where $\mathbf{U}$ is a matrix of a set of output orthonormal singular vectors of $\mathbf{X}^t$, $\mathbf{V}^t$ is a matrix of a set of input orthonormal singular vectors of $\mathbf{X}^t$ and $\mathbf{\Sigma}$ is a matrix of the singular values of $\mathbf{X}^t$.

Singular values of $\mathbf{X}^t$ are extracted from the diagonal matrix $\mathbf{\Sigma}$ and they are normalized with respect to the largest one. Finally, an estimation of the rank of matrix $\mathbf{X}^t$ is computed annotating the number of normalized singular values that surpass a predefined threshold.

The main idea of this shot boundary detector is that ideally the estimated rank of the matrix $\mathbf{X}^t$ will remain 1 within the same shot, will form a pulse shape along time when there is an hard transition and will from a triangle shape along time when there is a gradual transition. A visual representation of this concept in a gradual transition can be seen in Figure 2.4.

FIGURE 2.4: Example of rank evaluation in a gradual transition [23]

The proposed approach has been tested with groundtruth C in the beginnings, half-times and endings of the game with the same parameters depicted in [23]. The obtained results are shown in Table 2.3.

	Recall	Precision
SVD ranking analysis.	91.5%	84.4%

TABLE 2.3: Shot Boundary Detector Results for Gradual Transitions using a SVD ranking analysis.

It can be observed that the precision metric has been lowered considerably in contrast to the hard-cut transition detector. This is because in shots where the camera performs fast zoomings or pannings in order to change the focused player, the suggested method confuses the high motion as gradual shot boundaries.

To exemplify that shot detection is a complex task, in the following Figure 2.5 two specific examples of gradual transitions are shown. In the sequence A it can be seen that the color distribution difference among frames is low, fact that causes that the direct frame-by-frame histogram comparison is unable to detect the boundary. On the contrary, if the transition is evaluated with a sliding window filled with histograms of

frames, the boundary can be found. Nevertheless there are gradual transitions that only with color histograms cannot be detected, such as the one depicted in the sequence B, where the green predominates in the adjacent shots. However, this is not a common scenario.

FIGURE 2.5: Examples of Gradual Transitions: Sequence A and Sequence B

2.2 KeyFrame Selection

Keyframe extraction is an important task in many applications such as video editing and summarization. In this thesis this task has been implemented for the summarization of shots. Its main goal is to find the keyframes that will be subsequently analyzed by color descriptors, goal-post detectors and facial detectors.

A video sequence can be expressed as a combination of shots, each shot being a number of video segments consisting of a set of similar frames. Generally, the function of a keyframe selector is to choose a set of representative frames of each video segment. Yet, this is a complex task, because there is no objective and efficient method that allows such discrimination.

Many keyframe extraction techniques can be found in the literature [7][24][51]. One of the simplest methods is to select the first frame of each shot, but this frame does not have to adequately represent the rest of the shot. Other computationally efficient method is to sample the video stream directly and go uniformly capturing keyframes.

However, this generates random results and too many redundant keyframes. There are more intelligent techniques, such as those based on segmenting multiple shots into clusters and selecting as a keyframe the closest frame to the centroid of each cluster [24].

These segmentation techniques consist in extracting different characteristics of each frame to form a feature vector. Then clusters are created by measuring the similarities among frames taking into account the feature vectors. Finally frame distances inside the cluster are analyzed and keyframes are extracted.

There exist various types of characteristics to measure the distance between frames, such as those based on color, intensity of the movement, contours, textures or mathematical transform coefficients.

In this section two keyframe selection approaches are evaluated. First a method based on a classical color histogram approach [24] is introduced and later on a novel technique specially designed for the suggested framework combining motion and color is proposed.

2.2.1 Color Histogram based

This classical approach consists in segmenting shots in different clusters grouping those frames that share a high similarity in terms of color. It is based on the work of Y. Zhuang et al. [24] in which color histograms are used as main features to represent a frame. The chosen histograms are unidimensional and with 32 bins in order to achieve a good trade-off between the amount of information to compare and the computational cost used. In addition, only the blue component from an RGB frame is employed. This decision has been taken after performing several tests in the proposed scenario, soccer matches, and realizing that provides efficient and adequate results.

The implemented similarity measure to differentiate between histograms is the intersection, defined in equation (2.9). This is a low-cost computational measure in which the maximum similarity is the number of pixels of the image and a dissimilarity always produces an inferior result to this value.

$$sim(\mathbf{h}_1, \mathbf{h}_2) = \sum_{i=1}^{N} min(h_1(i), h_2(i)) \tag{2.9}$$

where N is the total number of bins and $\mathbf{h}_1$ and $\mathbf{h}_2$ are the color histograms.

The defined procedure for the keyframe selection is a non supervised algorithm that computes the distance from each frame to the existing clusters and if there's no cluster near the frame it creates a new one. Each cluster is represented with the average image

of the frames within the cluster and this image is updated when a new frame enters to the cluster. This iterative algorithm can be summarized in the following 8 steps:

1. Initialization: Create the first cluster c_1 and represent it by the first frame f_1. Number of clusters $numClusters$ equals 1.
2. Get the next frame f_i. If there are no more frames go to step 8.
3. Compute the similarity (2.9) between the frame f_i and the representative frames fc_k from the existing clusters $(c_1...c_k...c_{numClusters})$.
4. Determine which cluster is the closest to frame f_i defining maximum similarity as:

$$MaxSim = max_{k=1}^{numClusters} sim(\mathbf{h}_{f_i}, \mathbf{h}_{fc_k}) \tag{2.10}$$

5. Threshold the maximum similarity. If $MaxSim$ is below the thresold it means that f_i is not close enough to none of the clusters, then go to step 6. Otherwise, put f_i in the cluster with maximum similarity and go to step 7.
6. Increment the number of clusters, $numClusters = numClusters + 1$. Create a new cluster with the frame f_i as a representative. Go to step 2.
7. Update the representative frame fc_k from the cluster. Assuming that $fc_k^{'}$ is the previous representative frame and D the number of frames in the cluster, let us denote the new fc_k as (2.11). Go to step 2.

$$fc_k = \frac{D}{D+1} fc_k^{'} + \frac{1}{D+1} f_i \tag{2.11}$$

8. Finalization: For each cluster select as keyframe the image f_j closest to the representative frame of the cluster fc_k. This procedure can be formulated as:

$$keyframe_k = max_{j=1}^{D} sim(\mathbf{h}_{f_j}, \mathbf{h}_{fc_k}) \tag{2.12}$$

The generated results from a keyframe selector are difficult to evaluate objectively. The election process of a frame as a representative image from a video segment is a complex and very subjective task. Taking into account that the proposed clustering criteria is based on grouping images in terms of their color distribution, in the following figures the obtained results can be seen. In Figure 2.6 a video sequence formed by four shots is depicted, each shot being represented by a uniform two frame-per-second sampling. Selected keyframes from this sequence are represented in Figure 2.7.

It can be observed that in those shots where the color variation is low only a keyframe is selected (shot 2 and 3 from Figure 2.6). On the contrary, in those shots where there's

FIGURE 2.6: Video sequence of four shots showing two frames per second

FIGURE 2.7: Keyframes selected from the video sequence of Figure 2.6

a noticeable movement and the image varies considerably in terms of color, more than one keyframe are picked (shot 1 and 4 from Figure 2.6).

In general this algorithm provides acceptable results, but there's a specific characteristic in its nature that makes it unsuitable for our soccer scenario. When there's high motion in a shot several clusters are created and each one of them is represented by the averaged image of its frames. After that, the chosen keyframe is the most similar frame to this averaging of highly-changing frames. Therefore, in these occasions and due to the algorithm, keyframes tend to be quite blurry because the representative image tries to gather the information of as many frames as possible creating a blurred image. This could be a good strategy for video editing keyframes, but in our summarization framework, where keyframes' edge conditions are extremely important in order to detect goal-posts and faces, this is not a suitable approach.

The insights learned after implementing this methodology justify the proposal of a new strategy which is presented in the next subsection.

2.2.2 Hybrid Motion Activity and Color Histogram based

In order to solve the problems encountered in the previous approach, a novel and efficient iterative technique based on motion activity and color statistics was specially designed for the proposed framework.

As described in Chapter 1, the soccer video matches are compressed in MPEG-4 Simple Profile format with a resolution of 360x288 and a GOP structure IPPPPPPP (I7P). Bearing this in mind, all the motion vectors employed in this approach are extracted directly from the MPEG streams and motion activity is defined as the magnitude of these motion vectors.

The suggested keyframe selection procedure first segments the shots into multiple clusters taking into account high peaks of motion. Later on from each cluster it searches the predictive (P) frame with the lowest motion activity and selects as a candidate keyframe the nearest intra (I) frame. This is done to choose as keyframe an image as static as possible and with the highest quality. Finally the color resemblance among the candidate keyframes is assessed in order to eliminate highly similar frames.

Initially the algorithm processes all the motion compensation vectors within a shot and performs a full search analysis in order to find the predictive frame with the maximum motion activity, M_{max}. Then, M_{max} is compared against an adaptive threshold $T_1 = \alpha M_{median}$, defined as the median motion activity in the shot multiplied by a constant α larger than 1. The constant α specifies how much deviation from the median can be accepted as a maximum and the median measure is used in order to discard motion outliers. If $T_1 < M_{max}$, the shot is split into two sub-shots in the M_{max} time instant and the algorithm is applied again in both sub-shots. Otherwise, if $T_1 > M_{max}$, a search analysis is performed to find the minimum motion activity, M_{min}. Once M_{min} is found, the closest intra frame from the video sequence stream is labeled as a candidate keyframe. After analyzing all the frames in a shot or finding a maximum of 10 candidate keyframes the iterative algorithm stops.

The second stage measures the color similarity among the candidate keyframes. The keyframes that are closer in terms of color are discarded by doing a histogram comparison in the HSV colorspace using the Chi-Squared distance (2.1).

This algorithm is able to summarize each shot in keyframes depending on the intensity of action of the scene in a computationally efficient manner. Moreover it provides low-motion activity keyframes that usually yield a better performance in the person and goal-post detection phase.

2.2.3 Analyzing the keyframe selectors' results

In order to evaluate the results from the proposed keyframe selection procedures several examples of soccer scenarios are analyzed. In the following figures a set of extracted keyframes are compared. The left frames correspond to the keyframe selection based on color histograms and the right frames to the hybrid motion activity and color histogram approach.

In Figure 2.8 a typical example of a running player is depicted. It can clearly be seen that the right image provides a sharper picture where the player's face is more focused.

FIGURE 2.8: Keyframe comparison example 1: A running player

In Figure 2.9 there's a goal celebration. Left image shows a frame from a high-motion and jumping sequence and right frame represents an instant from a more static moment.

FIGURE 2.9: Keyframe comparison example 2: A celebration

Figure 2.10 depicts a zoom out operation of a goal-post. In this example the difference in sharpness is quite significant. The proposed goal-post detector would perform considerably better in the second picture due to the clearer definition of the edges.

FIGURE 2.10: Keyframe comparison example 3: A zoom out operation

On the other hand, focusing on the computational costs, the hybrid motion and color approach is quite faster than the color clustering method. This is mainly due to the fact that the segmentation of the second approach relies on the directly extracted motion compensation vectors from the MPEG stream while the color histograms from the first method need to be computed frame-by-frame.

To conclude, after having analyzed both keyframe selectors it can be stated that semantically speaking the validity of their keyframes is similar. However, the proposed keyframe extraction method based on motion activity and color histograms is faster and provides the sharpest keyframes. Therefore it is more suitable for the suggested framework where the keyframes need to be post-processed by goal-post and facial detectors.

In this chapter the soccer video sequence has been split into a minor temporal unit called shot. In order to detect shots boundaries several detectors have been suggested and a combination of a hard transition detector and a gradual transition detector has been proposed for the present framework. Later on each shot has been represented by a set of frames denoted as keyframes and the chosen method to do so has been a hybrid motion and color based keyframe selector.

Chapter 3

Description Tools

This chapter is devoted to the description tools that extract the set of audio-visual descriptors in the analysis bank (see Figure 1.1). These descriptors are computed for each one of the shots and stored in XML files. Once a shot is fully annotated all the XMLs belonging to it are grouped into a single file. In the next sections the approaches followed by the description tools are defined. In addition, the metrics employed to assess the performance of the approaches are described in Appendix A.

3.1 MPEG-7 Descriptors

In this framework the first step to describe the audio-visual soccer content is to annotate its shots and keyframes through the extraction of basic low-level descriptors. The selected low-level descriptors are specified by the MPEG-7 standard and their XML structure and syntax can be found in [25][26]. The goal of extracting MPEG-7 low-level descriptors is to gather basic audiovisual features that can be exploited a posteriori in order to generate new mid-level soccer-related descriptions. These low-level descriptors create an upper layer of information with a higher semantical meaning than the direct pixels and sound samples from the video and audio streams. MPEG-7 provides a set of visual descriptors that annotate color, motion, shape and texture from an image or a video sequence and a set of auditive descriptors that describe the temporal and frequency domains of an audio track. The shape and texture descriptors have been discarded in our framework because the employed techniques that use this type of information operate directly in the raw video stream. From the remaining descriptors, only a few from each category have been selected and the criteria employed is detailed at the end of this section.

In the following lines the importance of audio, motion and color descriptors for the soccer event detection process is justified and the software libraries used for their extraction are named.

Audio descriptors [25] play a crucial role in the summarization process of soccer sequences because the most important events tend to happen where the audio presents concrete features. For example goal occasions are usually followed by a sudden audio power increase and important events can be verified if a referee's whistle is spotted nearby. The MPEG-7 audio descriptors implemented in this framework have been extracted using the Java MPEG-7 Audio Encoder binaries [60].

Motion descriptors [26] describe the level of action in the scene (e.g. the peace in a panoramic view or the high intensity movement in a goal occasion) and at the same time they can provide information about camera operations such as pannings, tiltings, zooms-in, etc. These descriptors are efficiently computed exploiting the motion compensation information provided in the MPEG streams. This information is obtained analyzing the motion vectors from the Predicted (P) frames in our IPPPPPPP (I7P) GOP. The extraction of motion vectors has been done with the FFMPEG C library [61] and the motion descriptors have been implemented following the work in [62].

Color descriptors [26] in soccer events help in recognizing the type of view of the field (close-up shot, medium shot, long shot, the stands, etc.). The color descriptors have been calculated using the MPEG-7 Low Level Feature Extraction Library [40].

In order to select specific descriptors from each one of these three categories, several reasonings were followed: First we choose the most basic descriptors, the ones that we were certain that could be exploited in the soccer video field analysis (Dominant Color Descriptor, Color Layout Descriptor, Motion Activity Descriptor, Camera Motion Descriptor and Audio Power Descriptor). Then, we added the descriptors that we thought that could be employed in the future for advanced descriptions (Color Structure Descriptor, Audio Fundamental Frequency and Audio Spectrum Envelope). Finally, a decision was made to also extract the computationally inexpensive descriptors that were directly related with the previously computed ones (Audio WaveForm and Audio Spectrum Centroid). In the end only the basic descriptors were used and their corresponding descriptions can be found in the next sections. However, we strongly believe that the rest of the descriptors are also worth mentioning for future work. For instance the Color Structure Descriptor [26] captures the local spatial structure of the colours and the colour distribution of an image and could be used for classifying the viewpoints of the soccer pitch. Moreover, the Audio Fundamental Frequency and Audio Spectrum Envelope descriptors [25] could distinguish between different sounds of the match, such as the cheering, the referee's whistle or the voice over of the journalists.

To sum up, the following audio, motion and color MPEG-7 descriptors were selected and extracted employing the corresponding tools:

Audio domain: The Audio WaveForm, the Audio Power, the Audio Fundamental Frequency, the Audio Spectrum Envelope and the Audio Spectrum Centroid.

Motion domain: The Motion Activity Descriptor and the Camera Motion Descriptor.

Color domain: The Color Structure Descriptor, the Dominant Color Descriptor and the Color Layout Descriptor.

In the next sections the combination and comprehension of the low-level features from these MPEG-7 descriptors will aid in the creation of higher-level soccer descriptions.

3.2 Long Shot Detector

To automatically summarize a soccer match, the most significant highlights need to be identified. One way to simplify this process is to discard first the non-relevant occurrences. A shot that is usually of low interest in a soccer game is the one that offers a panoramic view of the pitch. This type of shot is called long-shot and is characterized by having green as the predominant color and a camera with slow movements. Examples of these viewpoints can be seen in Figure 3.1.

FIGURE 3.1: Examples of long-shots

Another type of shot where the green color can also dominate is the mid-shot. It is a viewpoint between a close-up and a panoramic shot in which players' entire body are shown. Unlike the long-shot, the amount of green leans to be lower and the colors are more compacted. In Figure 3.2 examples of mid-shots are depicted.

The goal of this section is to detect long-shots while clearly distinguishing them between mid-shots. The detection procedure consists in four stages and optionally an extra stage to reduce false positives. In the first stage the stands of the soccer pitch are trimmed, in the second the XMLs of the MPEG-7 descriptors are parsed, in the third it is found that the dominant color is green and lastly that this green color is evenly distributed.

FIGURE 3.2: Examples of mid-shots

Additionally, the optional final procedure to discard false positives is based on observing if there is a low number of camera operations in the shot.

The first step is to eliminate the non-relevant information to the analysis. This information are the non-green regions of the image and tends to correspond to the stands of the soccer pitch. Due to this peculiarity and after evaluating a set of football matches, this problem has been tackled by directly cutting the upper third part of the image, maximum space that usually takes this non-desired information. This technique sometimes excludes useful data but it is considered that the remaining pixels are discriminant enough to detect long-shots. An example of this trimming is represented in Figure 3.3.

FIGURE 3.3: Examples of the trimming stage

The next stage computes the chosen color and motion MPEG-7 descriptors, specifically, the Dominant Color Descriptor (DCD), the Color Layout Descriptor (CLD) and the Camera Motion Descriptor (CMD), which are briefly described below. The motion descriptors are computed from the temporal segment of each shot and the color descriptors from the selected keyframes. At the output of this stage XML files are generated whose metadata will be the input information for the posterior analysis.

To check that a shot is a long-shot, it is verified that green is the dominant color in a keyframe through the information provided by the DCD. This descriptor quantizes an image leaving only the eight most representative colors with their corresponding percentages [26]. Thus, this could be considered a filtering step where only those keyframes

with at least an 85% of green tonality pass. This tonality tries to gather all the possible green-grass like colors and is obtained by setting thresholds on the RGB color components.

The third stage assesses the green color homogeneity by analyzing the parameters from the CLD. This color descriptor reduces the input image into a tiny 8x8 YCrCb image and finds the Discrete Cosine Transform (DCT) coefficients. In the long-shot view, soccer players are represented with few pixels, therefore, after applying this down-sampling, they tend to be removed. This procedure provides a uniform green image in the long-shot scenario in contrast to the mid-shot case. Consequently the system analyzes the homogeneity in the green color by calculating the variance of the first 9 alternate coefficients (AC) of DCT chromas and comparing them with a threshold. The keyframes' variances that are below the threshold are marked as belonging to a long-shot view.

Finally, an optional filter to check if the candidate shots are static is proposed. This static filter relies on the CMD which is the descriptor in charge of annotating the amount of camera operations that appear in a video sequence and in our case, in a specific shot. The operations that we detect [62] with this descriptor are pannings, tiltings, zoomings and the fixed camera. Bearing this descriptor in mind and knowing that usually in long-shots there are no representative camera displacements, this stage consists in verifying if the fixed camera is the only operation registered in the CMD. This procedure efficiently reduces the number of false positives, even though at the same time a considerable quantity of true positive are removed. As a consequence, after performing some tests and observing that the filter was too restrictive, we decided to leave it as optional and disable it by default.

The proposed long-shot detector has been tested with keyframes coming from the the soccer matches of groundtruth A. The results from a total of 373 keyframes can be observed in the following Table 3.1.

	Recall	Precision
Default configuration	80.2%	96.3%
Static filter enabled	67%	100%

TABLE 3.1: Long-shot detection results

As can be seen the difference between the precision and the recall is significant. A higher precision has been preferred over a high recall. Being the objective of long-shot detection its posterior discarding, it has been assured that what is discarded is actually non-relevant. It is worth mentioning that activating the optional filter to exclusively detect statics long-shots increases the precision up to 100%, although the recall falls to

67%. On the other hand, non-detected long-shots have been those in which the camera focuses on the lower, far left or far right part of the pitch. In these cases the lack of green color causes that 85% of the pixels are not of green tonality, what produces that the third decision threshold is not surpassed. Examples of these non-detected shots are represented in Figure 3.4.

FIGURE 3.4: Examples of non-detected long-shots

3.3 Zoom Detector

The aim of this description tool is to detect events of interest from the viewpoint of the director of the match. These events reflect the moments when the director wanted to emphasize details of the video sequence and did so through the zoom-in camera operation.

There are many algorithms in the literature for zoom detection [41][42][43]. The most common methods involve computing the optical flow, fitting its motion vectors into a parametrical motion model and assessing the corresponding coefficients of the model. Unfortunately the majority of these steps include iterative techniques that require high computational costs. Alternative simpler methods are based on segmenting the image into different regions and correlating the motion of each region with a number of specific patterns [41]. Examples of these patterns can be found in Figure 3.5. The latter

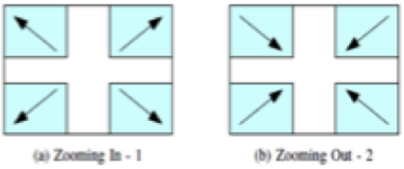

FIGURE 3.5: Examples of templates to detect zooms [41].

zoom detection algorithm works properly in events where the camera does not perform translation operations and assumes that the zoom focus is at the center of the image. However, in sports the camera is constantly moving and the zoom operation is suddenly

mixed with previous translation operations. This causes that not all motion vectors follow the same pattern and that the focus of the zoom is not always placed in the center. Because of these difficulties we have chosen the proposed procedure by Huang Jing-hua and Yang Yan-song[42] that estimates the center of focus without using predefined patterns. Their approach involves segmenting the motion field into several regions, computing a representative motion vector for each section and assessing if the direction of these vectors converge in one central focus.

Initially, a 23x18 motion vector field is extracted from the 360x288 MPEG video sequence and then motion vector outliers are removed using a median filter. This filter consists of a 3x3 mask that runs throughout the motion field, orders the 9 vertical and horizontal components of each vector and selects those that fall in the middle. A sample output of this filter can be seen in Figure 3.6 where a panning to the right is shown. Blue motion vectors denote 0 magnitude.

FIGURE 3.6: Left image: Original motion field. Right image: Filtered motion field.

The next step is to segment the motion field with the golden section spatial composition rule [42]. This rule suggests splitting the image into 3:5:3 both vertically and horizontally as shown in Figure 3.6. The importance of these divisions comes from the fact that normally the director of sport events tries to frame interesting details in the intersections of these regions.

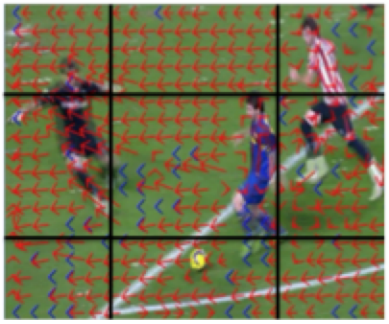

FIGURE 3.7: Golden section spatial composition rule

Once the motion field is segmented into the 9 golden sections, the 8 external regions from the image are analyzed. Firstly it is evaluated whether the information provided in the areas is of interest. To do so, the vectors' magnitude in each region are computed and it is checked if two thirds of them are different from zero. Then, in all areas that exceed the previous stage the average motion vector $\mathbf{v}_{avg}$ is calculated and the vector of the region that most resembles to it is sought with the equation given by (3.1). $\mathbf{v}_r$ is the representative vector of the r_{th} region and expresses the pointed direction by the movement of the area.

$$\mathbf{v}_r = \arg\min_{\mathbf{v}\epsilon\mathbf{S}_r} \sqrt{(\mathbf{v}-\mathbf{v}_{avg})^T(\mathbf{v}-\mathbf{v}_{avg})} \tag{3.1}$$

where $\mathbf{S}_r$ is the set of motion vectors $\mathbf{v}$ from the r_{th} region.

Later on every vector is extended into a straight line, passing through the source and destination points of every vector, and all crossing points caused by the intersection of pairs of lines are found. The purpose of this step is to estimate the zoom focus through a centroid formed by the intersection of points. This centroid is computed by averaging the crossing points positions. In Figure 3.8 a representation of this zoom focus estimation can be seen. Red arrows denote the position of the representative vectors and the blue circle specifies the center of the estimated focus.

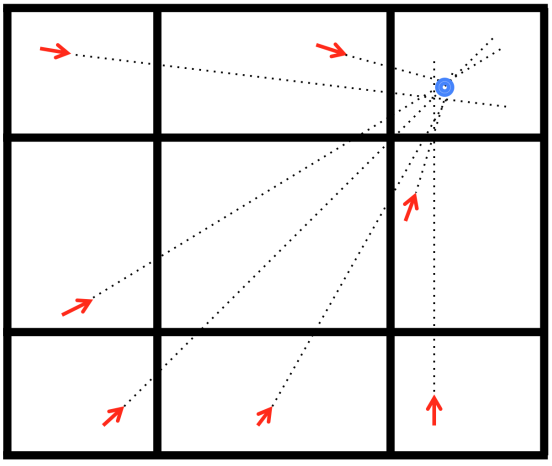

FIGURE 3.8: Zoom focus estimation.

The last step to detect the existence of a zoom consists in measuring the distance between the intersection points' position and the estimated focus' position. Ideally all lines coming from the representative vectors should converge in the center of the focus, thus the distance would be zero. However, due to all the performed estimations regarding the prediction of motion vectors and the zoom behavior, this distance is not null and

needs to be compared to a low-valued threshold. In our case, taking into account the 360x288 resolution of the video sequences, the selected threshold has been 0.4.

Finally, once a zoom has been detected, in order to identify if it's a zoom-in or a zoom-out, the direction of the representative vectors is assessed. This zoom detector has been tested in the groundtruth A containing 120 soccer zooms and the results are depicted in Table 3.2.

Recall	Precision
78.33%	88.68%

TABLE 3.2: Zoom detection results

Given that in our application a higher precision rate than a higher recall is preferred, the achieved results are considered satisfactory. Most of the non-detected zooms are found in the time instants in which apart from the zoom operation other movements occur in the video sequence. These movements are generally local motions caused by the displacement of players and significantly worsen the global motion analysis.

3.4 Goal Post Detector

In order to summarize sport events it is of great significance to know where the action is taking place. Specifically in sports where important events tend to occur near a concrete zone in the pitch such as soccer. Due to this, goal-post detection has been a hot area of research in the soccer summarization literature [53][54][55] and a description of it has been implemented in this framework.

Many goal-post detection approaches [53][54] consist in using the Hough transform in order to identify the vertical goal posts. Others employ morphological operations [55] using a line as a structuring element and detect the vertical and horizontal posts. Even though these are good approaches, being the goal-post viewpoints quite limited, we proposed a novel method to tackle this problem based on object modeling.

The proposed method relies on describing goal-posts using the Histograms of Oriented Gradients (HOG) [56] as features and training a latent support vector machine (LSVM) leveraging deformable part-based models [57].

3.4.1 Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) [56] has been a well-known feature in the past few years in object recognition. Initially it was developed for detecting pedestrians in static images but later on it was expanded to other fields. Its main idea is to divide the image in different regions, called cells, and a for each cell compute the histogram of the orientation of the gradients. After that, the combination of these histograms represents the descriptor.

HOG is a heavily engineered descriptor and to design it the authors tested a lot of parameters, in the following paragraph a brief summary of the default approach that yielded the best results is presented.

In order to compute this descriptor the gradients of an RGB image are extracted employing a simple derivative mask [-1 0 1] in the vertical and horizontal dimensions. Then each pixel in a 8x8 pixels cell votes for an edge orientation histogram channel and each vote is weighted by the pixel's gradient magnitude. The orientation bins are unsigned and evenly spaced over 0-180^o into 9 histogram channels. Afterwards, 2x2 cells are grouped into a larger region called block and each block is normalized using a modification of the L2-norm [56]. The aim of this local normalization of cells is to account for variations in illumination and contrast. Finally, being blocks 50% overlapped, the descriptor is defined as a detection window composed of 64 x 128 pixels, divided into 105 blocks, 7 blocks horizontally and 15 blocks vertically. The final HOG feature vector is composed of 7 blocks horizontally x 15 blocks vertically x 4 cells per block x 9 bins per histogram = 3780 values.

Once a large number of HOG descriptors are extracted from the same object, to create a model the authors propose training a linear SVM. Later on, to recognize an instance of this object, the matching procedure consists in a traditional detection window that scans the input image using different positions and scales [56]. Nevertheless, model creation and matching procedure for HOG descriptors have evolved and alternatives have appeared that increase the overall recognition performance and the computational costs. In this framework the proposed alternative is a deformable part-based model trained with a LSVM. Thus, in the following subsection the details of this approach are introduced.

3.4.2 Deformable part-based model

Object detection with deformable part-based models is a computer vision framework defined in [57] that tackles the problem of non-rigid deformations and multi-scale variations in objects. The framework describes an object with a star-structured part-based model defined by a root filter plus a set of parts filters and associated deformation models. These filters are learned with a latent support vector machine (LSVM) and their goal is to weight the HOG features and provide an object detection score.

The root filter is devoted to characterize the object as whole while the set of part filters specify important regions of the object. For instance in a person model the root filter consists in the person's body and example of part filters can be the head and the feet. In addition each part filter is associated with a deformation model which penalizes and measures the relative distance of the part filter to the root filter. All filters and their corresponding deformations are automatically trained with the LSVM. An example of these trained filters and deformations can be seen in the person model of Figure 3.9. This model [57] has been trained using 2416 positives images and 1218 negative images from the INRIA Person Dataset [56].

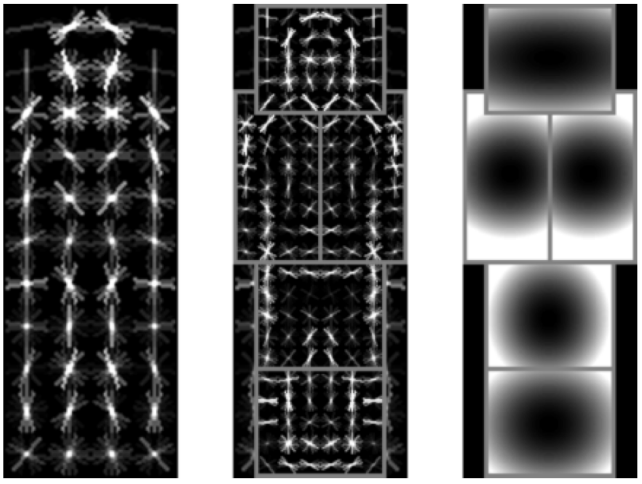

FIGURE 3.9: Deformable part-based model of a person [57]. Being the first image the root filter, the second the part filters and the third the deformation model.

These visual representations show the positive weights of the filters at different orientations and the penalizations of placing a part filter at different locations relative to the root filter.

This a multi-scale model in which the input image is scaled at different levels and a feature pyramid is built. Besides, the authors noticed that part filters are of paramount importance and that providing a higher resolution in them increased the performance of the object detection task. Consequently, being the root filter in a specific level in

the pyramid, the part filters are represented with twice the resolution to that level (see Figure 3.10).

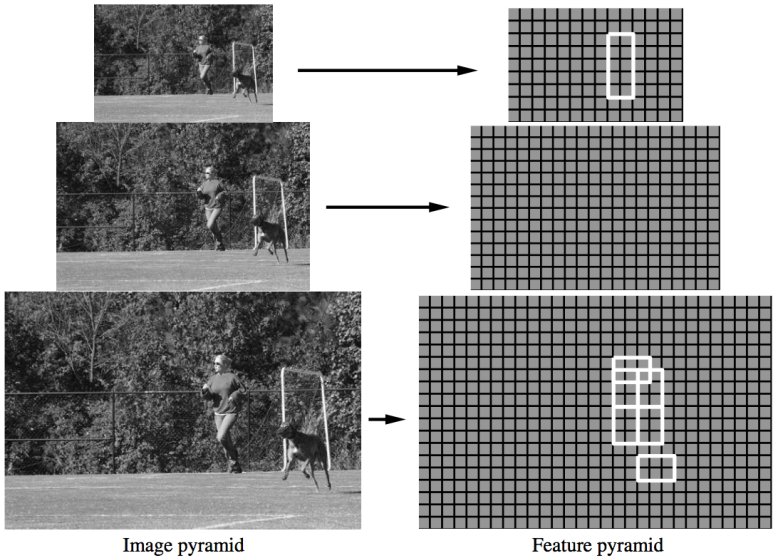

FIGURE 3.10: A feature pyramid and an instantiation of a person model within that pyramid [57]. The first image is in scale 1x, the second in scale 1.5x and the third in scale 2x.

To define the function to be trained by the LSVM, let $\mathbf{p} = (x, y, l)$ specify a pixel position (x, y) in the l_{th} pyramid level. Let $\mathbf{H}$ be a HOG feature pyramid matrix. Let $\mathbf{f}$ denote a filter vector and $\mathbf{d}$ a deformation cost vector, both vectors to be trained. Let i be the filter index being $i = 0$ the root filter and $i > 0$ the part filters. Then, let us give a score of an object hypothesis by the sum of filter scores minus the sum of the deformation scores as depicted in equation (3.2).

$$score(\mathbf{p}_0, ..., \mathbf{p}_n) = \sum_{i=0}^{n} \mathbf{f}_i^T \phi(\mathbf{H}, \mathbf{p}_i) - \sum_{i=1}^{n} \mathbf{d}_i^T \phi_d(dx_i, dy_i) \tag{3.2}$$

where $\phi(\mathbf{H}, \mathbf{p}_i)$ is a function that provides a feature vector at $\mathbf{p}$ position and

$$\phi_d(dx_i, dy_i) = (dx_i, dy_i, dx_i^2, dy_i^2) \tag{3.3}$$

is the function that provides the quadratic deformation features being dx_i and dy_i the displacement of the i_{th} part filter relative to the root filter.

The object detection score in each image position $\mathbf{p}$ consists in defining the score of each root filter location as the score given the best part filter placements (3.4). Therefore, the root filter is fixed at position $\mathbf{p}_0$ while the different part filters are evaluated at several locations $\mathbf{p}_i$.

$$score(\mathbf{p}_0) = \max_{\mathbf{p}_1,...,\mathbf{p}_n} score(\mathbf{p}_0, ..., \mathbf{p}_n) \tag{3.4}$$

A visual scheme of the overall object detection procedure using the deformable part-based models is represented at Figure 3.11.

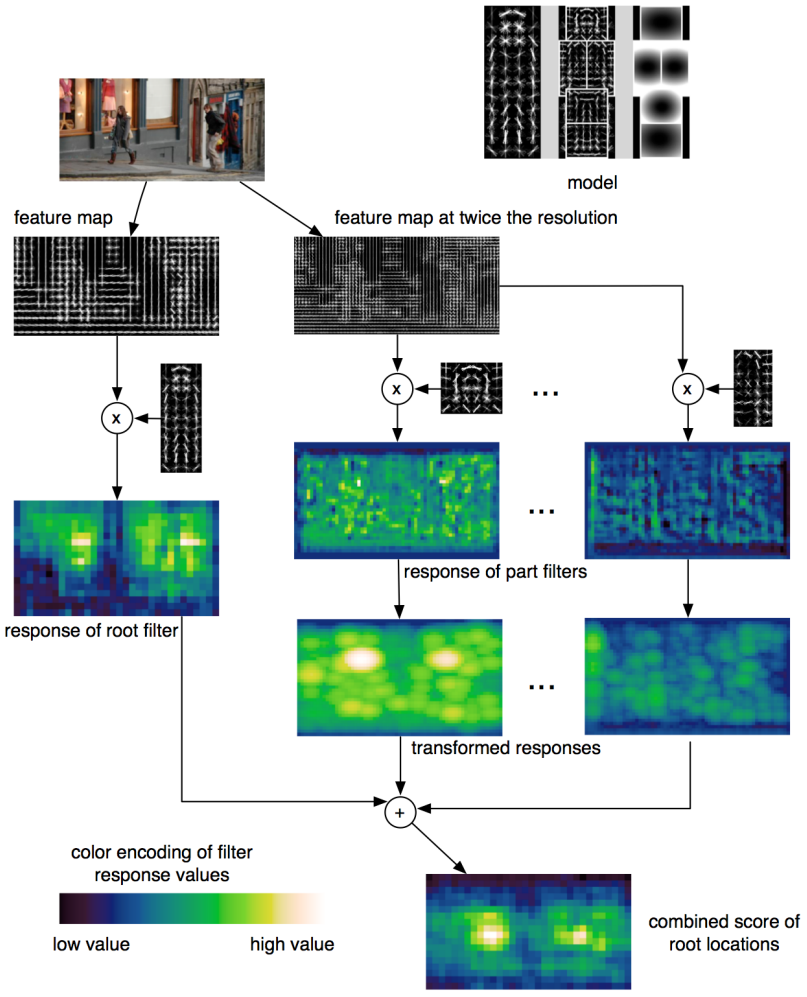

FIGURE 3.11: Overall object detection procedure using the deformable part-based models with a trained person model [57].

Lastly, in order to train the LSVM let us rewrite the equation (3.2) into a shorter equation (3.5). The LSVM differs from the traditional SVM because extra latent parameters need to be learned. In our case this latent information are the set of positions and bounding boxes dimensions of the part filters in the object model denoted by $\mathbf{z}$.

$$score(\mathbf{z}) = \mathbf{w}^T \Phi(\mathbf{x}, \mathbf{z}) \tag{3.5}$$

where $\mathbf{z}$ is the latent information; $\mathbf{w}$ are the concatenation of the root filter, part filters and deformation models; $\mathbf{x}$ is the input image; and $\Phi(\mathbf{x}, \mathbf{z})$ is the function that provides the HOG and deformation features.

With the equation (3.5) defined, the LSVM is trained [57] and the deformable part-based model is created. In the following subsection this procedure has been applied to obtain a goal-post model.

3.4.3 Training a goal-post deformable part-based model

To train the deformable part-based model a database consisting in 500 goal-post images was created from groundtruth A. The images were selected and classified taking into account its right, left and frontal orientations. In the following Figure 3.12 examples of these orientations can be seen.

FIGURE 3.12: Examples of goal-post orientations.

The database was split into a training set and a evaluation set, each one composed equally of 250 images. Furthermore, the images belonging to the training set were segmented using the LabelMe [58] annotation tool and a total of 250 goal-posts were precisely labeled following its boundary edges (see Figure 3.13).

FIGURE 3.13: Edge segmentation with LabelMe.

Later on to create the models the training software provided in [57] was used and the LabelMe annotation outputs were adapted to it. Finally the goal-post deformation part-based models were generated and their representations can be seen in Figure 3.14.

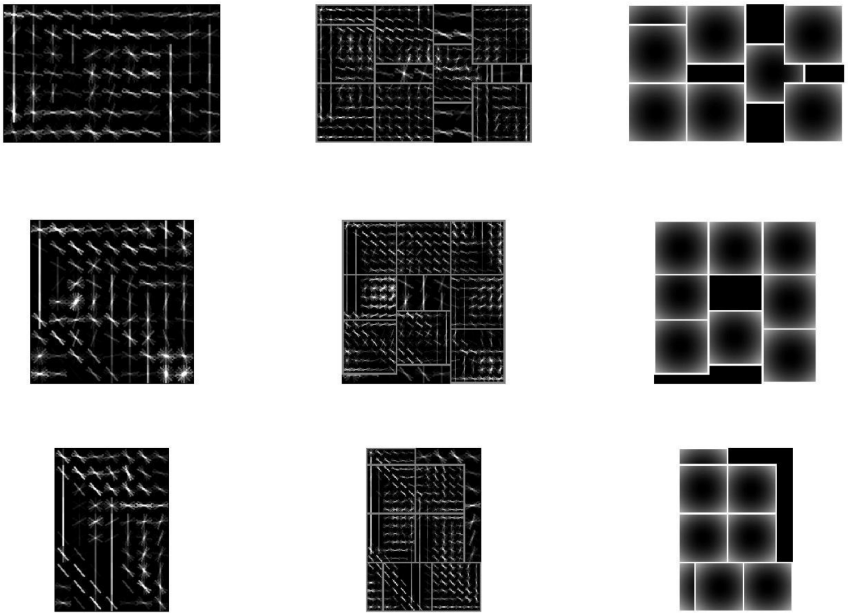

FIGURE 3.14: Goal-post deformation part-based models.

The three models denote clearly the majority of viewpoints from our goal-posts training set. Leaving only those points of view in which not enough examples were selected. These models were tested against the evaluation set where 200 samples belonged to same soccer matches as the training set and 50 were from a completely different match. The goal-post detection results are presented in the following Table 3.3.

Recall	Precision
71.77%	99.33%

TABLE 3.3: Goal-post detection results

The results are believed to be extremely good given the fact that the training database is significantly small. Examples of correctly detected goal-posts can be seen in Figure 3.15, being the red boxes the root filters and the blue boxes the part filters.

FIGURE 3.15: Correctly detected goal-posts.

Almost all the non-detected goal posts are frontal ones, from backwards and from far away distance. This is because the database does not contain a significant number

of these viewpoints and a model for them has not been created. Figure 3.16 shows representative non-detected samples and the only false positive detected.

FIGURE 3.16: Non-detected goal-posts (left) and false positive (right)

We strongly believe that if the current database were increased with more samples from the non-detected goal-post points of view (far away distance, frontals and backwards) the recall metric would improve and even better results could be achieved.

3.5 Overview of other tools

The following tools are going to be described in more generic way because even though they are included in the framework and the highlight generator leverages them, they have not been directly implemented by the author of this thesis.

3.5.1 Persons Detector

The detection of soccer players in close-up shots is a very important feature of the highlights summary generator system. However, in order to develop an efficient system of soccer player's detection, some constraints derived from the soccer scenarios need to be taken into account. In particular, images with low resolution and/or high luminance variation, faces with different scale factors, rotations, poses and/or occlusions, profile faces tilted up to 90^o as well as high angles face shots can be found in the soccer scenario. Although results on sports players detection have been reported [27][28][29], and there are a number of general face detection schemes [30][31], there is a need to develop more robust schemes taking into account these soccer-based constraints. In this context a robust persons descriptor based on a face detection approach has been implemented targeted to detect players in a soccer video sequence. The approach is based on the well-known Viola and Jones AdaBoost [29] approach face detector. The results of this filter are confronted with an skin filter that has been included to decrease the number of false alarms. Let us note, that the person descriptor can also be applied to detect events where persons are involved. Figure 3.17 shows some examples of close-ups detected with the proposed approach. These examples include a profile case, detections in the bench and different poses and high angle shots.

FIGURE 3.17: Face detections in soccer. Profile case (a). Detections in the bench and public examples (b). Different poses and high angle shots (c), (d) and (e).

The general overview of the system is shown in Figure 3.18, where several classifiers are used for the face detection stage using the approaches presented by Viola and Jones [29] and implemented in [34].

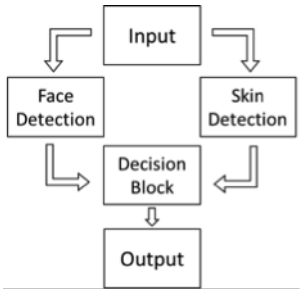

FIGURE 3.18: General overview of the persons descriptor system

A specific trained classifier to detect head and shoulders has also been used [35] to increase the detection rate. The skin filter is derived from the method proposed in [32] but using the RGB and HSV color subspaces instead of the Normalized Color Coordinates (NCC). Subsequently a simple condition adapted from [33] is defined for the HSV subspace to refine the detection. Finally, a decision block to reduce the false alarm level is applied. The proposed approach takes into account all the possible ethnic groups.

The recall and precision of the proposed face detection approach using groundtruth B present values of 92.8% and 89.45% respectively. In this case, to check the robustness of the approach, the test has been performed with a database formed by 550 images with multiple faces from the second and third Spanish soccer division, which present a worse scenario (less cameras, worse illumination) than first division since the TV production for these categories have a different production process. The good results found in these adverse and demanding scenarios proof the validity of the approach to be used in easier 1st division matches.

3.5.2 Replay Detector

The most interesting events in a soccer game are usually presented through replays from different scene viewpoints or in slow motion. Hence, replay detection approaches can offer reliable descriptors in the highlights summarization process.

The TV production style in soccer media usually identifies the beginning and ending of a replay using two identical logos. Therefore, the strategy to detect a video sequence soccer replay can be based on logo detection [36][37][38][39] and is generally implemented in four stages: searching for video frames that are candidates to contain a logo, detection of the logo pattern used in the soccer media, matching of the logo pattern along the soccer video and pairing the detected logos to identify the replays.

The proposed replay detector is presented in Figure 3.19 and it is based on [39], that has been modified to improve the false detection rate.

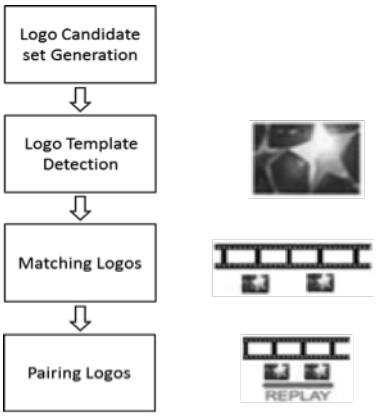

FIGURE 3.19: Replay detector scheme

The algorithm employed in the first step uses the luminance difference between frames to detect the peaks along the video frames that characterize a logo presence. This method is suitable for hard logo transitions; however, as it presents a low performance for gradual transitions, the algorithm in [39] has been modified with the aid of a shot boundary detector [23]. Then a k-means clustering is applied on the luminance and variance of the frame images in order to detect the logo template. Once the logo templates have been detected, the logos and the candidate images are converted to gray scale, and a pixel-by-pixel subtraction is performed. The sum of all the differences gives an idea about their similarity. Finally, a threshold is applied to detect the correct matches that present the lowest difference values.

The proposed approach has been tested in a database formed by different logo patterns from groundtruth B with a total number of logos to be detected of 507 and 226 replays. Assuming the logo template is known, the recall and precision in the logo detection

process is of 100% and 99.6% respectively. Independently, in the pairing logo stage the recall is 99.12% and 100% for the precision, which means some replays are discarded due to logos not being correctly paired. A replay is identified if and only if its beginning and ending logos are detected. In TV production styles that only use single logos, the pairing logos stage is not applied.

3.5.3 Whistle Detector

In many sports, such as soccer, the detection of the referee's whistle provides highly valuable information to detect events of interest. Therefore, reliable and robust whistle detection is a key objective in the design of methodologies for automatic sport highlighting. Whistle detection has been considered extensively in the literature and different strategies have been proposed [45][46][47][48][49]. Most of these proposals are based on analyzing the spectral content of the signal and detecting the maximum energy in the whistle's frequency band. Although these methods produce acceptable results in some scenarios, the number of false detections may be significant in recordings with too much cheering noise nearby the whistle's frequency band. The method proposed in this thesis tries to improve these results by further exploiting the spectral characteristics of professional whistles.

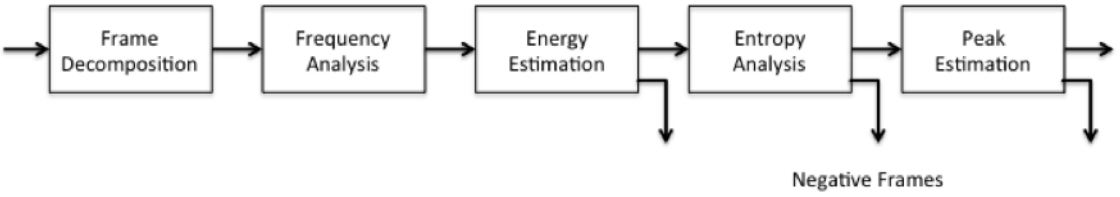

FIGURE 3.20: Whistle detector scheme

The first block segments the audio signal onto frames. In our database, a sampling frequency of 48 kHz, and a frame size of 4,800 samples have provided accurate results for the next stages.

The frequency analysis block computes a set of Discrete Fourier Transform (DFT) samples in the band of interest for every audio frame. That band is selected between 3.5 kHz and 4.5 kHz as broadly include the frequencies produced by professional whistles. The reduced set of samples of the DFT is computed using the Goertzel algorithm [50].

The next stage estimates the energy contained in the interest band of an audio frame. Once this energy is estimated, a threshold is applied to detect the whistle. Although there exists high correlation between energy peaks and whistles, in some difficult scenarios thresholding produce an unacceptable number of false alarms due to the presence

of noise in the band of interest. Then, the last two stages of the whistle detector system try to reduce the number of false detections by exploding the fact that the whistle spectrum is made of 3 tones at very close frequencies. However, as these frequencies are not stable and may vary slightly with the specific excitation, the direction of blowing and the whistle model, in order to discriminate the tonal vs non-tonal nature of the audio frame, an entropy-based stage is introduced.

The approach consists in considering a normalized version of spectrum samples in the interest band as being samples of a probability density function. When the entropy is computed on these samples a number that indicates the spread of the samples in the frequency band will be obtained. High numbers indicate wide spread while small numbers indicate that the energy is concentrated on a few, high probable energy samples. This entropy-based concept is defined as:

$$H(m) = \sum_{k=K1}^{K2} p_m(k) \log_2 p_m(k) \qquad with \;\; p_m(k) = \frac{\left|X_m(k)^2\right|}{\sum_{r=K1}^{K2} |X_m(r)^2|} \tag{3.6}$$

where $K1$ and $K2$ represent the DFT samples at the limits of the interest region and m represents the frame time index.

Applying a threshold on this entropy value permits to discard all those audio frames where the energy is spread over the interest band. The tonal audio frames that do not exceed that threshold will be further processed by the final stage. Entropy analysis rejects some of the false positives in the example. Finally the last stage consists in discarding some sounds that are sometimes confused with the whistle by selecting the total number of peaks that exceed a threshold. Only audio frames with 2 or 3 peaks are finally validated.

In order to test the performance of the whistle detector a database has been created by manually annotating the complete recordings from groundtruth A. The annotation is performed using both video and audio tracks. Video track gives a helpful context required to discern referee's whistles among other sounds such as vuvuzelas, horns, supporter's whistles, etc. Aside from this database a test signal has also been generated. This test signal is made up of a selection of referee's whistles annotated in the database in conjunction with other especially difficult sounds, usually inducing false positives, which have been encountered in these recordings. The total length of the test signal is around 60 s and contains 10 referee's whistles, which are represented in Figure 3.21. Figure 3.22 represents the energy estimate. The result of applying the entropy function to the test signal is represented in 3.23 where it can be verified that the entropy decreases in the frames that correspond to whistles. The results obtained when applying the algorithm to the complete database show a recall of 93% and a precision of 88%.

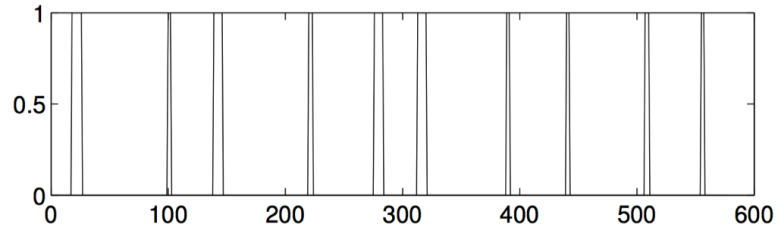

FIGURE 3.21: Groundtruth

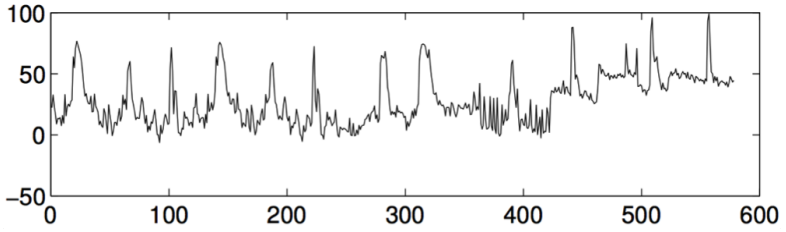

FIGURE 3.22: Energy in the interest band

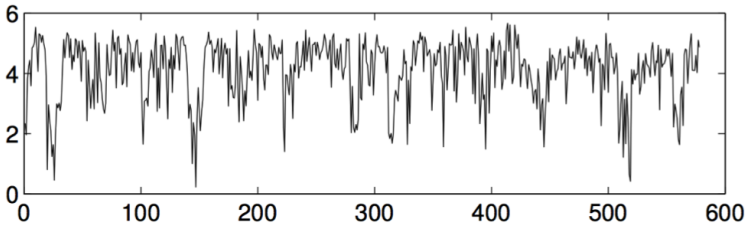

FIGURE 3.23: Entropy estimation of the spectrum

3.5.4 Inter and Intra shot-based audio power detectors

Temporal audio changes play a crucial role in the detection of relevant soccer events. Hence, to spot important instants in the audio track, three detectors that measure audio power variations have been extracted:

• The A.Power.H and A.Power.VH detectors represent peak levels of audio power within a shot, where H stands for high and VH for very high. Logical binary values are associated to those shots whose maximum audio power is over 95% and 97%, respectively, in reference to the maximum audio power value of the entire audio soccer track.

• A.IntraInc.50 and A.IntraInc.100 represent audio power increments within a shot. The low-level audio power descriptors are averaged for every second. The first detector is used to represent logical true values when the audio power in these averaged intervals is increased in 50% whereas the second detector represents increments of 100

• A.InterInc.50 and A.InterInc.100 are the same as A.IntraInc.50 and A.IntraInc.100 but they refer to average audio power increments between contiguous shots.

These basic low-level descriptors will provide helpful and insight information to generate the video soccer match summarization when combined with other descriptors.

Chapter 4

Highlights Generator

The goal of this chapter is to define the tool that provides the video sequence summary of soccer highlights. First the architecture of the highlights generator is introduced along with an example of a configuration file and an output file of the system. Later on the results of detecting a concrete event, the goal, are presented and finally an analysis of the output summaries produced by the highlight generator system is done.

4.1 Highlights Generator Architecture

Once all the low-level and mid-level audio-visual descriptors have been extracted, the tool in charge of selecting the most important shots to include in the soccer summary is the highlights generator. The architecture of the system is shown in Figure 4.1 and consists of a filter bank and an event detector. The filter bank scores each shot using a set of elementary and advanced filters and the event detector generates the final summary by assessing the previous scores and a set of summary options defined by the user.

The filter bank is composed of elementary and advanced filters. Elementary filters are defined as filters that detect or classify events in each shot by using features directly provided by the low-level and mid-level audio-visual descriptors detailed in Chapter 3. Filters that fall into this category are: the long shot detector, the zoom detector, the whistle detector, the replay detector, the persons detector, the high motion detector, the goal-post detector, the audio intra power shot detector, the audio inter power shot detector, the long duration classifier, the medium duration classifier, the short duration classifier and the very short duration classifier. Each detector can be specified in five temporal time instants indicating whether the event has been detected two shots before, in the previous shot, in the actual shot, in the next shot or in two shots ahead. Duration

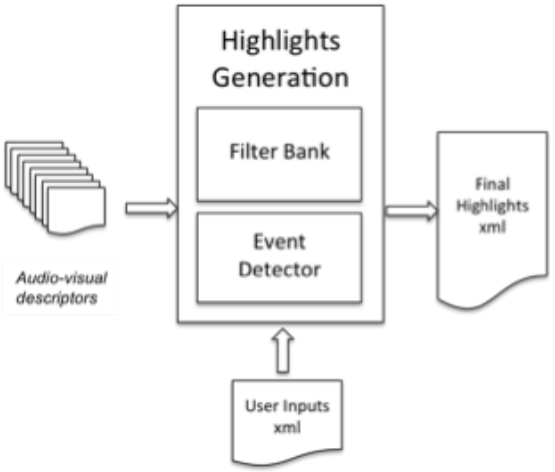

FIGURE 4.1: Highlight generator architecture diagram.

classifiers categorize the length of a shot. The long duration classifier annotates shots that last more than 20 seconds, the medium duration classifier shots between 10 and 20 seconds, the short classifier shots between 10 and 5 seconds and the very short classifier the shots shorter than 5 seconds.

The objective of the advanced filters is to provide a higher-level description of a shot by linearly combining elementary filters. To do so, each elementary filter is understood as a boolean function F, which is weighted by a coefficient w. These w coefficients can take positive or negative values in order to benefit or penalize the outputs of elementary filters. The result of an advanced filter operation is a local score L and their aggregation gives the global score G of that shot as defined in (4.1).

$$G = \sum_{j=1}^{M} L_j = \sum_{j=1}^{M} \sum_{i=1}^{N} w_{i,j} F_i \tag{4.1}$$

where M and N are the total number of advanced and elementary filters respectively.

Advanced filters can be interpreted as event detectors that may be used to identify goals, faults, important moments, goal-celebrations, etc. In our results, the weights of the elementary filters are selected manually based on previous knowledge about production and edition techniques. Alternatively, the weights could be acquired by training a linear classifier, but due to the different styles in broadcast soccer video generation and the enormous effort of creating a valid groundtruth for such highly semantic machine learning task, this option has been discarded.

A graphical matrix-based representation of the elementary and advanced filters concept is shown in Figure 4.2. Where each row denotes an advanced filter that combines elementary filter columns and green and red colors specify positive and negative weights

respectively. In addition, the results column provides the local scores of each advanced filter being its last row the global score depicted in orange color. As can be seen an advanced filter may be composed of only one elementary filter (A1, A2, A4) or more than one elementary filter (A3).

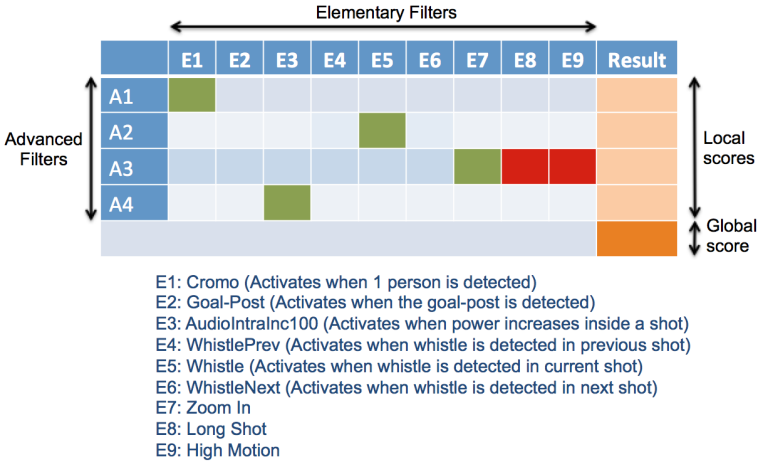

FIGURE 4.2: A matrix-based representation of the elementary and advanced filters concept.

Once all the shots are scored, the event detector stage is initiated. Here the aim is to choose the final shots that will produce the video sequence of soccer highlights. In order to do so this stage takes into account the user preferences about the summary. The user can specify in advance the length of the summary and also the percentage of appearance of each one of the advanced filters. As an example, the user could select a summary of 3 minutes where 50% of the time is dedicated to goals, 30% to important game moments, 10% to goal-celebrations and 10% to faults. The shot selection algorithm gathers the shots with the highest local scores for each one of the advanced filters, sorts them by their global scores and finally selects a number of shots till the desired duration is achieved. One problem that commonly arises is that there are too few shots for a specific advanced filter. In this case the system equally distributes the missing duration within the remaining advanced filters.

Also, a description is added to each shot specifying the events captured by advanced filters. An event is detected when all the elementary filters with positive weights in an advanced filter are triggered. An exception to the previous rule occurs when different temporal instances of the same detector are included in an advanced filter. In this case only one detector of all the temporal instances needs to be activated for the event to be annotated. For example, in the goal detector a replay is looked for in the next shot and also two shots ahead, the goal of this exception specifies that the detection of only one replay is enough.

In the following Figure 4.3 an example of an XML with a filter bank configuration is depicted. In this summary the desired duration is 10 minutes and the events to be scored by advanced filters are zooms in, replays, important moments, goals or occasions, whistles and shots with people in a close-up view (in soccer terminology this term is usually referred to as "cromo"). The percentage of appearance of every event is denoted by the tag *summary_percentage* and in Figure 4.3 it can be noticed that there are two filters with a 0 percentage of appearance, the audio intra increase 100 and the duration score. The sole purpose of these filters is to increase/decrease the global score of shots and in this case, to benefit the shots with short duration and an important audio increase. An analysis of the design and the results of an advanced filter is done in the next subsection.

```
<?xml version="1.0" encoding="UTF-8"?>
<FilterBank>
        <Id>GeneralSoccerSummary</Id>
        <Duration>10</Duration>
        <Filter id="1" description="ZoomIN" summary_percentage="10">
                <ZoomIN>3</ZoomIN>
                <HighMotion>-3</HighMotion>
                <LongShot>-3</LongShot>
        </Filter>
        <Filter id="2" description="Replay" summary_percentage="20">
                <Replay>5</Replay>
        </Filter>
        <Filter id="3" description="ImportantMoment" summary_percentage="20">
                <LongShot>-7</LongShot>
                <APowerVeryHighNext>2</APowerVeryHighNext>
                <AudioIntraInc100>3</AudioIntraInc100>
                <CromoNext>2</CromoNext>
        </Filter>
        <Filter id="4" description="GoalOrOccasion" summary_percentage="30">
                <LongShot>-7</LongShot>
                <APowerVeryHighNext>2</APowerVeryHighNext>
                <AudioIntraInc100>3</AudioIntraInc100>
                <GoalPost>2</GoalPost>
                <WhistleNext>3</WhistleNext>
                <CromoNext>2</CromoNext>
                <CromoNext2>2</CromoNext2>
                <ReplayNext>5</ReplayNext>
                <ReplayNext2>5</ReplayNext2>
        </Filter>
        <Filter id="5" description="Whistle" summary_percentage="10">
                <Whistle>3</Whistle>
        </Filter>
        <Filter id="6" description="Cromo" summary_percentage="10">
                <Cromo>1</Cromo>
        </Filter>
        <Filter id="7" description="AudioIntraInc100" summary_percentage="0">
                :AudioIntraInc100>2</AudioIntraInc100>
        </Filter>
        <Filter id="8" description="DurationScore" summary_percentage="0">
                <DurationShort>3</DurationShort>
                <DurationMedium>1</DurationMedium>
                <DurationLong>-3</DurationLong>
                <DurationVeryLong>-10</DurationVeryLong>
        </Filter>
</FilterBank>
```

FIGURE 4.3: Example of a filter bank configuration xml.

The output of the highlight generator is an XML file that contains the time codes of selected shots. The shots are sorted following the original order of their appearance in the match and each one of them has a description tag that specifies the events that have been detected in it. In the file there is also metadata about the source match of

the summary, the template used in the filter bank and the length of the summary. An example of the structure of the XML is shown in the next Figure 4.4.

```
<?xml version="1.0" encoding="UTF-8"?>
<SummaryMetadata>
        <IdSourceMatch>8579546-BarçaMalaga.mp4</IdSourceMatch>
        <IdFilterBankTemplate>GeneralSoccerSummary</IdFilterBankTemplate>
        <Duration>10</Duration>
</SummaryMetadata>
<SummaryTimeCodes>
        <Shot id="110" tcIn="00:11:15:18" tcOut="00:11:16:22" eventDescription="Cromo|Whistle"/>
        <Shot id="112" tcIn="00:11:18:03" tcOut="00:11:19:05" eventDescription="Replay"/>
        <Shot id="114" tcIn="00:11:22:12" tcOut="00:11:26:13" eventDescription="ZoomIN"/>
        <Shot id="118" tcIn="00:11:35:02" tcOut="00:11:47:11" eventDescription="Cromo"/>
```

FIGURE 4.4: Example of the highlight generator output.

4.2 Highlights Generator Experiments

In order to test the validity of the audiovisual framework for automatic soccer highlights generation, two types of experiments have been done and analyzed. Firstly the design of an advanced filter, the goal or ocasion detector filter, is discussed and then its performance is assessed. Secondly, two automatic soccer summaries are generated and the shots included in them are manually categorized in order to analyze what type of events are being captured and how they are distributed. In addition, five individuals have watched the generated summaries and expose their points of view about the limitations of the produced video sequences. The database used to evaluate the performance of the experiments are the five soccer matches from groundtruth A. Similar results have been found for other groundtruth sequences.

An example of an advanced filter is a goal-scoring detector. Goal scoring is one of the most interesting and relevant actions in soccer. However, its highly semantic level makes its identification a very difficult task, specially if only low-level descriptors are available. Therefore, in order to tackle this problem we combined low-level and mid-level descriptors and also assumed that we could not distinguish between a goal occasion and an actual goal.

As it can be seen in the advanced filter number 4 in Figure 4.3, we define a goal or an occasion as an abrupt increment of the audio power (A.Intra) in a non long-shot (LongShot) where the goal-post (GoalPost) appears, followed in the near shots by a loudly soccer player's celebration (Cromo and A.Power), a replay detection (Replay) and a whistle (Whistle). As one could imagine, it is quite improbable that all the ingredients of this pattern are meticulously followed all the time. Nevertheless, we observed that very often even though one or two elementary filters are not triggered, the others alone are able to fulfill the task. This reasoning was the main purpose of choosing this type of scoring-based method. On the other hand, due to this flexibility

an important problem arises when not all the positive elementary filters have been activated in an advanced filter and the system wants to label what type of event has been detected. This is an issue that we leave for future work and the following results evaluate the *GoalOrOccasion* advanced filter alone measuring exclusively the goals that can be detected with it. Bearing this in mind, Table 4.1 contains the detected goals for the soccer matches presented in groundtruth A employing the proposed advanced filter in summaries of 10 minutes length.

Goals	Match 1	Match 2	Match 3	Match 4	Match 5
Total Number of Goals	2	4	3	6	3
Total Detected Goals	1	4	3	4	1
Goals Not Detected	1	0	0	2	2

TABLE 4.1: Goal Filter Performance

The filter is able to detect satisfactorily over the 70% of the total amount of goals of the five soccer matches analyzed. However its performance presents some unpredictable limitations due to changes in TV production style or in external uncontrolled factors such as for instance the public that influences the audio power producing peaks and/or significant increments. It is also important to highlight that soccer match 4 and 5 present a higher number of goals not detected and the main reason stems from the fact that these sequences do not contain replays.

The second experiment in this section assesses the highlight generator output by analyzing the type of events that are selected in two automatic soccer summaries. Additionally five individuals express their opinions about the main differences between the proposed generated video sequences and human-made summaries. Bearing these tasks in mind, the chosen length for generating each video summary has been 15 minutes and the advanced filters and percentages employed have been the same as the ones depicted in Figure 4.3. These parameters aim to imitate as much as possible manually generated summaries of long duration made by Catalan TV editors.

For each summary the resulting shots have been classified by a person in five groups: goals, offense/defense, offside-faults, persons and replays. The goals category contains the shots where a goal or a goal opportunity occurs; opportunity of goal is any offensive play that ends close to the goal or due to a goalkeeper intervention. Offense/defense represents the offensive and defensive actions that do not entail a goal or a goal opportunity. Offside/faults groups faults, offsides and corners. Persons represent the shots that contain public or persons as soccer players, coaches or other team members. Replays are shots that contains replays of faults, offside, goals and other interesting events.

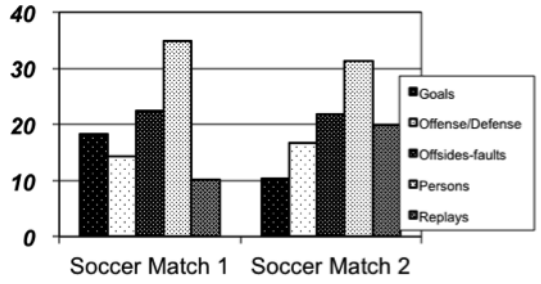

FIGURE 4.5: Classified shot distribution of two soccer matches

The resulting distributions of the classified shots are presented in Figure 4.5. Each bar represents the actual percentage of appearance of every category in an automatic summary. The graph shows that the suggested filterbank configuration is able to gather events of all the categories satisfactorily. Additionally, it can be observed that the persons and offense/defense classes have the highest bars and that their percentages are similar in both summaries. This reflexes the tendency of having an abundance of these types of shots in every soccer match and the easiness for the system to capture these events with the proposed configuration parameters. On the other hand, offside/faults, goals and replays are categories that are less likely to occur and heavily rely on the soccer match activity, factors that make that their percentages vary significantly depending on the match. Also, it is worth noticing that the person's category is the most popular shot type and surpasses the employed configuration that specifies a 10 percentage of appearance in the *Cromo* advanced filter. This is because others filters, such as the *ZoomIn* and *Whistle*, can also capture a shot with a person and in addition, it has to be reminded that if there's not enough shots to fulfill the duration of a specific filter, typical situation for the goal or the replay filters, the percentage of the remaining filters increases.

On the other hand, although the objective of this project was not to produce soccer summaries ready to the end-user, five individuals external to the project were asked to watch the generated video sequences to test the subjective quality of the automatic soccer highlights extracted by the proposed framework. The selected personnel are used to watch TV broadcasting soccer summaries and their task consisted in identifying the main limitations of our automatically produced summaries with respect to manually generated ones. The following statements are entirely subjective ideas that all of them agreed on:

- Although most important events are correctly detected not enough time is dedicated to some of them. For example, the precise time instant where a goal is scored is included in the summary, but not how the players reach to that event.

- Watching an automatic generated summary provides a good overview of some of important events of the match. However, as it can be imagined, it is unknown to the end-user if other very interesting occasions are missing, specially the goals.
- The tempo of the proposed summaries change very unnaturally compared to human-made summaries. For instance, a set of high motion shots can be found followed directly by a still shot, which translates into a non smooth human-pleasing transition.
- Two important premises fixed by the authors are that an interesting occasion occurs when the audience cheers and there is not a long-shot. Though this pattern is followed most of times some special play can still be missed.

Solving the issues depicted in the previous paragraph is clearly out of the scope of this thesis, but we believe it is good to take them into account for future work. Overall the group is very satisfied with the results obtained with the automatically generated soccer highlights and specially with the proposed audiovisual framework. More in depth conclusions can be found in the next chapter.

To sum up, in this chapter the architecture of system in charge of selecting the shots to include in the summary of soccer highlights is presented. The first section is devoted to detail the flexibility of the proposed scoring-based method and provides examples of filters that capture soccer events. Afterwards the assumptions and results of the suggested goal-scoring detector are depicted. Lastly, an analysis of the automatic soccer summaries generated by the system is done by categorizing the resulting shots, assessing their distributions and expressing their limitations with respect to manually generated summaries.

Chapter 5

Conclusions

In this section a brief summary of each one of the chapters is presented followed by the conclusions and future work.

This thesis presents a novel framework for generating automatic video highlights of soccer broadcasting video sequences. The proposed approach consists on segmenting the video sequence into shots and scoring them using a set of low-level and mid-level descriptors.

Chapter 1 provides an overview of the global system in which the structure and the followed strategy of the framework is presented.

Chapter 2 introduces the visual segmentation. Several shot boundary detectors and keyframe selectors are assessed and later on a combination of approaches is suggested. The proposed approach spots shot boundaries employing a hard cut transition detector in the action of the game and a cross dissolve detector in the beginning, half-time and final parts of the game. The hard cut detector is based on a traditional histogram frame-by-frame comparison and achieves a recall of 95.2% and a precision of 98.8% and the gradual detector consists on evaluating the rank of a matrix filled with histograms through the SVD and obtains a recall of 91.5% and a precision of 84.4%. Afterwards keyframes are selected taking into account codification, motion and color characteristics providing non-similar sharp intra frames.

Although the precision of the cross dissolve detector is low and extra efforts could be put in the future work to improve this part, our group believes that this should not be a priority task because gradual boundary detection is a difficult field, the number of these type of transitions in soccer is quite low and their apparitions are mostly in non-relevant moments of the match. On the contrary, an important addition for the visual segmentation module would be a sub-segmentation procedure for shots of long

duration. There are occasions when the number of cameras that record a game are very limited - for instance in a match from third division - and the soccer video sequence is composed of a reduced set of long duration shots. In cases like these an algorithm that analyzed the inside of a shot and were able to split it in a smart way would aid in great measure in the summarization task. Apart from this, the author is very satisfied with the results achieved with the keyframe selector, especially for the obtained image quality in the frames, the computational speed and the algorithm's ability to choose the appropriate number of keyframes to summarize a shot.

Chapter 3 presents the description tools used in the analysis bank. In this stage a set of low-level and mid-level descriptors are extracted from the shots and keyframes. Low-level descriptors describe audio, color and motion through MPEG-7 standards and mid-level descriptors annotate goal posts, persons, whistles, zooms, long-shots, replays and inter and intra shot-based audio power detectors. The goal post detector relies on a trained Histogram of Gradients (HOG) object-based descriptor and obtains a recall of 71.8% and a precision of 99.3%. The persons descriptor is based on adaboost-based detectors and skin filters and presents values of 92.8% in recall and 89.45% in precision. Zoom operations are extracted parametrizing the motion vectors and finding a common focal point and the detection results are of 78.33% in recall and 88.68% in precision. The whistle detector depends on the spectral energy, the entropy of the spectrum and a number of peaks in an interest frequency region and shows a recall of 93% and a precision of 88%. Long-shots are detected combining the MPEG-7's color layout descriptor and dominant color descriptor and the detection reaches an 80.2% in recall and 96.3% in precision. The replay detector finds the logos used in production that define the beginning and/or ending of a replay and assuming that the logo template is known, the recall and precision are of 100% and 99.6% respectively. Finally, the inter and intra shot-based audio power detectors capture temporal audio changes measuring audio power variations.

Taking into account the recall and precision metrics, the descriptors that the author would firstly refine would be: The goal-post detector, the zoom detector and the long-shot detector. Initially a descriptor that could be improved easily is the goal-post detector. In the current framework the precision of the proposed object detection approach is extremely high, however the recall can be refined. This is because only three points of views of a goal-post have been modelized. Generating a larger database would allow the creation of more models and then the recall metric would rise. Differently, the zoom detector would greatly benefit from a preciser and cleaner motion field than the one directly extracted from the MPEG stream. Therefore, although the computational costs would significantly increase, a hierarchical block matching algorithm could be recommended in order to calculate this previous information. Lastly, the false negatives of

the proposed long-shot detector mainly occur when besides the soccer pitch other outside regions appear in the image. Thus, substituting the cropping stage with a refined image segmentation technique where the grass could be correctly distinguished from the stands and other non-desired regions would also tremendously improve the recall number. For this task the author proposes a histogram back projection approach in order to satisfactorily identify the green region of the pitch.

Chapter 4 is devoted to the highlights generation module, which linearly combines and weights each descriptor by taking into account end-user preferences. In this module elementary and advanced filters are defined. Elementary filters classify events using low-level and mid-level descriptors and advanced filters provide a higher-level description by combining elementary filters. At the output of this procedure the proposed framework is tested with five soccer matches for goal-scoring detection and two custom soccer highlights are analyzed and subjectively evaluated.

From the point of view of the members of the AudioVisual Technology Group (GTAV) the achieved results are quite satisfactory. Generating an automatic soccer highlights summary of 15 minutes length provides a video sequence that retains most of the important events of a match and this can fulfill the main objective of the project of speeding the summarization task of Catalan TV journalists. Nevertheless, as it has been stated, these are not already prepared summaries for broadcasting purposes, specially because the loss of not detecting a goal is not acceptable in a professional environment and moreover because there's a lot of work to do for making these video sequences as visually pleasing as hand-made ones.

In the author's opinion future work should be focused on three well differentiated directions. First, the actual groundtruth of soccer matches could be increased altogether with their annotations, secondly the refinement and inclusion of more mid-level descriptions could be taken into account in order to provide more flexibility for the design of the event detectors and finally, the weights in the advanced filters could be trained with machine learning algorithms to improve the event detection accuracy. In the following lines the previous statements are justified and explained in more detail.

First of fall, enormous efforts should be made in order to create a large and smartly categorized groundtruth. After watching numerous soccer matches and talking with several people from the world of soccer editing, it has been noticed that there are different production styles in the soccer broadcasting field because there's no written rules about how soccer should be broadcasted. Therefore, it can be stated that each director has its own methods which translate into possible different patterns for representing every play. This means that, for instance, having a unique advanced filter for detecting a goal is not enough if a good generalization is desired. Consequently, having a huge number

of manually labeled soccer matches as examples is extremely important. Moreover, the variety of annotations that categorize the video sequences are also of major relevance because apart from regular labels, like the division number, the name of the soccer teams, the tags specifying where the goals are in the time-line, etc.; other non typical descriptions such as the editor of the match, the name of the stadium, the name of the referee, the hour when the match starts, etc.; could also provide really useful insights. For example, the specification of the person who is directing the match could facilitate the job of a machine learning algorithm in the identification of production style patterns. Additionally, in Catalonia the director of a soccer match tends to be directly related with the importance of the event, as famous directors are in charge of producing the most watched soccer matches such as the well-knowns Barça-Madrid. Therefore, extra information such as whether the match is going to be crowded or not, thus if the audio descriptors can play a relevant role, could also be inferred from this type of description. Differently, the name of the stadium could inform about the state of the grass and the number of cameras; the hour where the match starts could provide information about the lighting conditions, etc. In 1st division soccer metadata could be obtained from international video sport feeders or text-based broadcasting websites where all the matches are completely manually described. In addition, sport federations such as the UEFA and the FEF provide official proceedings indicating the most relevant events in a match and some sport news websites include automatic text-based soccer summaries; this could add an extra upper layer of metadata in which the level of interest of each soccer event is specified. Although this variety of annotations only matters if a large training database were built, the author believes that is good to take them into account for future work in a long-term view.

On the other hand, the adjustment and the addition of new mid-level descriptors could allow the creation of more accurate advanced filters. One descriptor that could be useful for the *Cromo* advanced filter is a t-shirt classifier. *Cromos* play an important role in soccer summaries because people enjoy watching the reaction of players, however, in the current framework although we include a large quantity of this type of shots, we cannot be certain if the majority of *Cromos* belong to the players from one team or if they are equally distributed. This descriptor would allow to choose the distribution of *Cromos* for each team and in an apparently easy manner, because no t-shirt recognition should be required, only the classification of the players chest in two major classes. Another very convenient annotation would be a ball trajectory descriptor. Still, to perform such a precise task a high resolution video would be needed. Based on this premise, the suggested approach would be to use a tracking algorithm that followed a model of the white color and the circle shape of the ball. As a tracking method the author proposes a Sampling Importance Resampling (SIR) particle filter where no linear assumptions

would be made about the movement. Other interesting descriptors would be a referee detector, a referee's card identifier, a penalty-box detector, medium pitch circle detector, etc.

Lastly, with the creation of the large and rich groundtruth previously depicted, the specification of the weights of the advanced filters could be done by a machine learning algorithm. This would increase the event detection accuracy and aid in the creation of new advanced filters. It is important to mention that the current framework structure could be kept by training with a linear classifier. Independently, apart from learning the weights of the advanced filters, in the author's opinion no training should be required for deciding which events enter in the final summary. This is a very subjective task and it is believed that manually picking the important events and its corresponding lengths using the proposed framework is a good strategy. Improvements in the highlight generator module should consist in specifying if an event should be summarized only with one shot, as it is now, or whether the surrounding shots may be included as well. This previous feature would allow a better representation of a goal and other events in the final summary.

Aside from improving the existing framework, future work could also consist in adapting the current tools for other applications. For instance, developing a user interface software where all the descriptions highlighted by the elementary filters and the events captured by the advanced filters could be viewed and searched easily along with the video sequence. This could aid TV journalists in order to generate their manual summaries and also could benefit trainers in their strategic analysis of soccer matches. Moreover, in a non-professional environment where the loss of some events could be tolerated, the current system could be employed in order to generate automatic summaries ready to the end-user. This could be the case of elementary school soccer competitions where usually a parent records the whole match and then wants to generate a short length summary for storing purposes.

Bibliography

[1] http://www.cenitbuscamedia.es

[2] Hu W, Xie N, Li L, Zeng X, Maybank S (2011) A Survey on visual content-based video indexing and retrieval. IEEE Transactions on Systems, Man, and Cybernetics. Part C: Applications and Reviews Vol. 41, no. 1, pp. 797-819.

[3] Datta R, Joshi D, Li J, Wang J Z (2008) Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, Vol. 40, no. 2, Article 5.

[4] Lew M S, Sebe N, Djeraba C, and Jain R (2006) Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computer and Communication Applications, Vol. 2, no. 1, pp. 1–19.

[5] Schoeffmann K, Hopfgartner F, Marques O, Boeszoermenyi L, Jose J M (2010) Video browsing interfaces and applications: A review. SPIE Review, Vol. 1, no. 1, pp. 018004.1–018004.35.

[6] Ji Z, Su Y, Qian R, Ma J (2010) Surveillance video summarization based on moving object detection and trajectory extraction. (ICSPS 2010) 2nd International Conference on Signal Processing Systems (ICSPS 2010), pp. v2-250-v2-253, July 5-7, 2010, Dalian, China.

[7] Truong B T, Venkatesh S, (2007) Video abstraction: A systematic review and classification. ACM Transactions on Multimedia Computer, and Communication Applications, Vol. 3, no. 1, art. 3, pp. 1–37.

[8] Xiong Z, Zhou S, Tian Q, Rui Y, Huang T S (2006) Semantic retrieval of video: Review of research on video retrieval in meetings, movies and broadcast news, and sports. Signal Processing Magazine, Vol. 23, no. 2, pp. 18 - 27.

[9] Ying L, Shih-Hung L, Chia-Hung Y, Kuo C-C J (2006) Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. Signal Processing Magazine, Vol. 23, no. 2, pp. 79 - 89.

[10] Ren W, Singh S, Singh M, and Zhu Y S (2009) State-of-the-art on spatio temporal information-based video retrieval. Pattern Recognition, Vol. 42, no. 2, pp. 267–282.

[11] Lavee G, Rivlin E, Rudzsky M (2009) Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews, Vol. 39, no. 5, pp. 489–504.

[12] Lotfi E, Pourreza H R (2007) Event detection and automatic summarization in soccer video. 4th Iranian Conference on Machine Vision and Image Processing (MVIP07), 2007 Mashhad, Iran.

[13] Xie L, Chang S F, Divakaran A, Sun H (2002) Structure analysis of soccer video with hidden Markov models. International Conference on Acoustic, Speech and Signal Processing, (ICASSP 2002) Vol. 4, pp. 4096–4099, May 13-17, Orlando, USA.

[14] Xu P, Xie L, Chang S F, Divakaran A, Vitro A, Sun H (2001) Algorithms and system for segmentation and structure analysis in soccer video. Proceedings of IEEE Conference Multimedia and Expo, (ICME 2001), pp. 928–931, August 22-25, Tokyo, Japan.

[15] Tabii Y, Oulad Haj Thami R (2009) A new method for soccer video summarizing based on shot detection, classification and finite state machine. 5th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications (SETIT 2009), pp. 7-11, March 22-26, Hammamet, Tunisia.

[16] Ekin A, Tekalp M A, Mehrotra R (2003) Automatic soccer video analysis and summarization. IEEE Transactions on Image Processing, Vol. 12, no. 7, pp. 796-807, July.

[17] Sadlier D A, O'Connor N E (2005) Event detection in field sports video using audio-visual features and a support vector machine. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, no. 10, pp. 1225-1233.

[18] Zawbaa H M, El-Bendary N, Ella Hassanien A, Kim T (2012) Event detection based approach for soccer video summarization using machine learning. International Journal of Multimedia and Ubiquitous Engineering, Vol. 7, no. 2, pp 63-80, April.

[19] Tavassolipour M, Karimian M, Kasaei S (2014) Event Detection and Summarization in Soccer Videos Using Bayesian Network and Copula. IEEE Transactions on Circuits and Systems for Video Technology, Vol.24, no.2, pp.291,304, February.

[20] Hanjalic A, Shot-boundary detection: Unraveled and resolved? (2002) IEEE Transactions on Circuits Systems and Video Technologies, Vol. 12, no. 4, pp. 90–105, April.

[21] Yuan J, Wang H, Xiao L, Zheng W, Li J, Lin F, Zhang B (2007) A formal study of shot boundary detection. IEEE Transactions on Circuits and Systems for Video Technology, Vol.17, no. 2, pp. 168-186, February.

[22] Smeaton A F, Over P, Doherty A R (2010) Video shot boundary detection: Seven years of TRECVid activity. Computer Vision and Image Understanding, Vol. 114, no. 14, pp. 411–418.

[23] Abd-Almageed W (2008) Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing. 15th IEEE International Conference on Image Processing, ICIP 2008, San Diego, California, USA, pp. 3200-3203, October 12-15.

[24] Zhuang Y, Rui Y, Huang T S, Mehrotra S (1998) Adaptive key frame extraction using unsupervised clustering. IEEE International Conference in Image Processing, pp. 283–287, October 4-7, Chicago, USA.

[25] Kim H G, Moreau N, Sikora T (2005) Low-Level descriptors, in MPEG-7 audio and beyond: audio content indexing and retrieval, 1st edition, New York: Wiley, ch. 2, pp. 13-59.

[26] Sikora T (2001) The MPEG-7 visual standard for content description-an overview. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, no. 6, pp. 696-702, June.

[27] Ballan L, Bertini M, Del Bimbo A, Nunziati W (2007) Automatic detection and Recognition of players in soccer videos. Proceedings of the International Conference on Visual Information Systems (VISUAL), Shanghai, China, June 28-29.

[28] Mahmood Z, Ali T, Khattak S (2012) Automatic player detection and recognition in images using AdaBoost. (IBCAST), 9th International Bhurban Conference on Applied Sciences and Technology, pp. 64-69, Jan 9-12, Islamabad, Pakistan.

[29] Viola P, Jones M (2004) Robust real-time face detection. International Journal of Computer Vision Vol. 57 no 2, pp. 137–154.

[30] Yang M H, Kriegman D J, Ahuja N (2002) Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Inteligence, Vol. 24, Issue 1, pp 34-58.

[31] Zhang C, Zhang Z (2010) A Survey of recent advances in cace detection. Microsoft Research Technical Report, MSR-TR-2010-66.

[32] Soriano M, Huovinen S, Martinkauppi B, Laaksonen M (2000) Using the skin locus to cope with changing illumination conditions in color-based face tracking. IEEE Nordic Signal Processing Symposium, pp. 383-386.

[33] Phung S L, Bouzerdoum A, Chai D (2005) Skin segmentation using color pixel classification: analysis and comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, no. 1, pp. 148-154.

[34] http://opencv.willowgarage.com/wiki/.

[35] http://mozart.dis.ulpgc.es/Gias/modesto.html.

[36] Zhonghua D, Jun D, Qingming H, Shuqiang J (2007) Replay detection based on semi-automatic logo template. Fourth International Conference on Image and Graphics, Chengdu, China August 22-24.

[37] Xiaofeng T, Hanqing L, Qingshan L, Hongliang J (2004) Replay detection in broadcast sports videos. Third International Conference on Image and Graphics, Hong-Kong, China, December18-20.

[38] Jinjun W, Engsiong C, Changsheng X (2005) Replay detection: scene transition structure analysis. International Conference on Acoustic, Speech and Signal Processing, (ICASSP 2005) March18-23, Philadelphia, USA.

[39] Wei X, Yang Y (2011) A robust replay detection algorithm for soccer video. IEEE Signal Processing Letters, Vol.18, Issue 9, pp. 509-512, September.

[40] Bastan M, Cam H, Gudukbay U, Ulusoy O (2010) BilVideo-7: An MPEG-7 compatible video indexing and retrieval system. IEEE Transactions on Multimedia, Vol. 17, no. 3, pp. 62-73, July-September.

[41] Nguyen N T, Laurendeau D, Branzan-Albu A (2010) A robust method for camera motion estimation in movies based on optical flows. The 6th International Conference on Information Technology and Applications, ICITA 2009, pp. 228-238, November.

[42] Jing-Hua H, Yan-Song Y (2008) Effective approach to camera zoom detection based on visual attention. 9th International Conference on Signal Processing, ICSP 2008, pp. 985-988, October 26-29.

[43] Superiori L, Rupp M (2009) Detection of pan and zoom in soccer sequences based on H.264/AVC motion information. 10th Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2009, pp. 41-44, May 6-8.

[44] Eldib M Y, Zaid, B, El-Zahar M, El-Saban M (2009) Soccer video summarization using enhanced logo detection. 16th IEEE International Conference on Image Processing (ICIP 2009), pp. 4345-4348, El Cairo, Egypt, Nov. 7-10.

[45] Tjondronegoro D, Phoebe Y P, Pham B (2003) Sports video summarization using highlights and play-breaks. 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 201-208. New York. 2003.

[46] Karthirvel P, Sabarimalai M, Soman P, (2011) Automated referee whistle detection for extraction of highlights from sports video. International Journal of Computer Applications. Vol 12 – no. 11, pp. 16-21.

[47] Zhang D, Ellis D, (2001) Detecting sound events in basketball video archive. Electrical Engineering Department of Columbia University. Speech and Audio Processing class Project Report.

[48] Bonarini A, Lavatelli D, Matteucci M A (2006) Composite system for real-time robust whistle recognition. RoboCup 2005: Robot Soccer World Cup IX. Springer-Verlag Lecture Notes in Computer Science Vol. 4020, pp. 130-141.

[49] Mellinger D K, Martin S W, Morrissey R P, Thomas L, Yosco J J (2011) A method for detecting whistles, moans and other frequency contour sounds. Journal Acoustic Society of America Vol 129, June.

[50] Oppenheim A V, Schafer R W (2010) Discrete-Time Signal Processing. Third Edition. Prentice Hall Signal Processing Series.

[51] Omidyeganeh, M., Ghaemmaghami, S., Shirmohammadi, S. (2011) Video Keyframe Analysis Using a Segment-Based Statistical Metric in a Visually Sensitive Parametric Space. IEEE Transactions on Image Processing. Vol.20. No.10. pp.2730-2737. Oct.

[52] M. N. Do, "Directional multiresolution image representations," Ph.D. dissertation, Dept. Commun. Syst., Swiss Federal Inst. of Technol.

[53] Yow D, Yeo B.-L, Yeung M, and Liu B (1995) Analysis and presentation of soccer highlights from digital video. Comp. Vision in Proc. Asian Conf. (ACCV).

[54] Zawbaa H.M, El-Bendary N, ella Hassanien A, Abraham A (2011) SVM-based soccer video summarization system, Third World Congress on Nature and Biologically Inspired Computing (NaBIC), pp.7,11, 19-21

[55] Nisha J, Santanu C, Roy S.D, Mukherjee P, Seal K, Talluri K (2008) A Novel Learning-Based Framework for Detecting Interesting Events in Soccer Videos, Sixth

Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP. , pp.119,125, 16-19

[56] Dalal, N, Triggs, B, (2005) Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, vol.1, pp.886,893, vol. 1, 25-25

[57] Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2010) Object Detection with Discriminatively Trained Part Based Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9.

[58] Russell B. C, Torralba A, Murphy K. P, Freeman W. T (2008) LabelMe: a database and web-based tool for image annotation, International Journal of Computer Vision, pp.157-173, vol. 77, no. 1-3.

[59] DeBrunner V , Kadiyala M (1999) Effect of Wavelet Bases in Texture Classification Using a Tree-Structured Wavelet Transform, IEEE Transactions on Image Processing, 33(4):1292–1296, August.

[60] http://mpeg7audioenc.sourceforge.net/

[61] https://www.ffmpeg.org

[62] Raventós A (2011) Descriptors de moviment per l'anàlisi del contingut audiovisual, Treball de Final de Carrera, UPC-BarcelonaTech, http://upcommons.upc.edu/pfc/handle/2099.1/12917

Appendix A

Detection performance metrics

The majority of the presented tools for the automatic soccer highlights generation task are based on detection and classification procedures. Due to this, in order to assess the performance of the tools' approaches, the *Recall* and *Precision* metrics related to the pattern recognition field have been employed along with manually annotated databases.

Labeled databases are also called groundtruths and they mainly consist in a set of manually categorized samples. In our case, for each detection tool a specific groundtruth is generated being its samples binary classified into *True* or *False* categories (see Figure A.1). In this classification scenario the meaning of these labels can be defined as:

- True: The annotated sample from the groundtruth has a desired feature and should be detected by the classifier.
- False: The annotated sample from the groundtruth does not own a desired feature and should be discarded by the classifier.

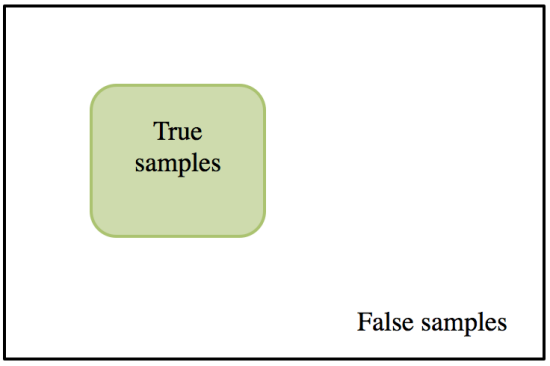

FIGURE A.1: Groundtruth classification.

Typically a groundtruth is divided into two sets, one set for training the classification algorithm and another set for testing the algorithm's performance. The proportion of

samples in each set can be selected and our approach has been splitting the groundtruth equally.

Once a classifier has been trained, it automatically categorizes unlabeled samples. The output categories of this procedure are the *Positive* and *Negative* labels depicted in Figure A.2. The significance of these classes is described in the following lines:

- Positive: An automatically detected sample by the classifier. The classifier's prediction is that this sample belongs to the *True* groundtruh category.
- Negative: An automatically non-detected sample by the classifier. The classifier's prediction is that this sample belongs to the *False* groundtruh category.

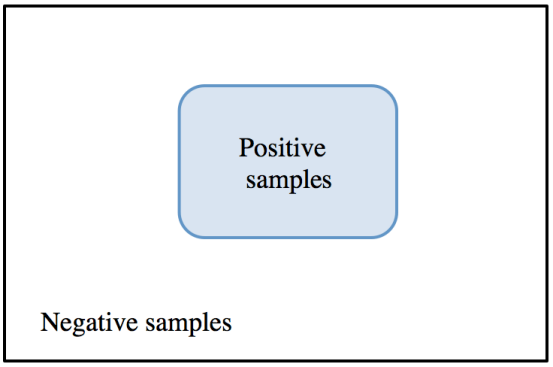

FIGURE A.2: Classifier classification.

To analyze the classifier performance the *True*, *False*, *Positive* and *Negative* classes are combined as can be seen in Figure A.3. This assessment of groundtruth and automatic classifications allows the creation of important concepts in the classification terminology such as the *True Positives* and the *False Positives*. Each one of them expressed as:

- True Positive: An automatically and correctly detected sample by the classifier. The number of *True Positives* can be computed as the intersection between *True* and *Positive* samples.
- False Positive: An automatically and wrongly detected sample by the classifier. The number of *False Positives* can be computed as the *Positive* samples minus the *True Positives*.

The remaining terms are the *True Negatives* and the *False Negatives*. However to understand these concepts the previous *True* and *False* concepts from the groundtruth labels need to be redefined. In this case *True* and *False* refers to whether the prediction made by the classifier is correct or not. Therefore the definition of the remaining terms is:

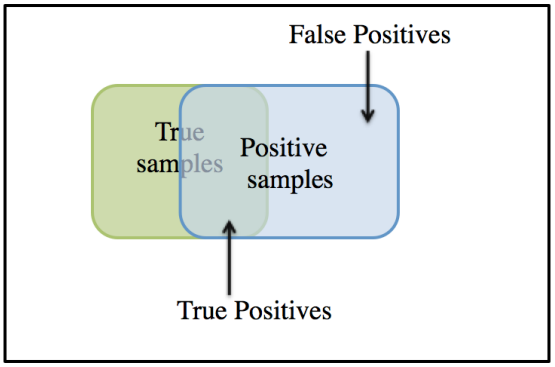

FIGURE A.3: Combining groundtruth and automatic classifications.

- True Negative: An automatically and correctly non-detected sample by the classifier.
- False Negative: An automatically and wrongly non-detected sample by the classifier.

Finally, being all these terms defined, the *Recall* and *Precision* metrics can be introduced. The *Recall* is called the sensitivity of the classifier and refers to the fraction of *True* samples that have been correctly detected (A.1). On the contrary, the *Precision* measures the fraction of automatically detected samples that have been correctly detected (A.2).

$$Recall = \frac{True\ Positives\ samples}{True\ samples} \tag{A.1}$$

$$Precision = \frac{True\ Positives\ samples}{Positive\ samples} \tag{A.2}$$

All the detectors presented in this thesis have been evaluated employing these *Recall* and *Precision* metrics.