A Deep Architecture for Multimodal Summarization of Soccer Games

Melissa Sanabria, Sherly, Frédéric Precioso
{sanabria,sherly,precioso}@i3s.unice.fr
Université Côte d'Azur, CNRS, I3S
Sophia Antipolis, France

Thomas Menguy
thomas@wildmoka.com
Wildmoka
Sophia Antipolis, France

ABSTRACT

The massive growth of sports videos, specially in soccer, has resulted in a need for the automatic generation of summaries, where the objective is not only to show the most important actions of the match but also to elicit as much emotion as the ones bring upon by human editors. State-of-the-art methods on video summarization mostly rely on video processing, however this is not an optimal approach for long videos such as soccer matches. In this paper we propose a multimodal approach to automatically generate summaries of soccer match videos that consider both event and audio features. The event features get a shorter and better representation of the match, and the audio helps detect the excitement generated by the game. Our method consists of three consecutive stages: Proposals, Summarization and Content Refinement. The first one generates summary proposals, using Multiple Instance Learning to deal with the similarity between the events inside the summary and the rest of the match. The Summarization stage uses event and audio features as input of a hierarchical Recurrent Neural Network to decide which proposals should indeed be in the summary. And the last stage, takes advantage of the visual content to create the final summary. The results show that our approach outperforms by a large margin not only the video processing methods but also methods that use event and audio features.

CCS CONCEPTS

• **Computing methodologies → Video summarization**; **Neural networks**; *Computer vision*; *Supervised learning by classification.*

KEYWORDS

Sports Summarization, Video Summarization, Multimodal analysis, Multiple Instance Learning, Long Short-Term Memory

ACM Reference Format:
Melissa Sanabria, Sherly, Frédéric Precioso and Thomas Menguy. 2019. A Deep Architecture for Multimodal Summarization of Soccer Games. In *2nd International Workshop on Multimedia Content Analysis in Sports (MMSports '19), October 25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3347318.3355524

MMSports '19, October 25, 2019, Nice, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6911-4/19/10…$15.00
https://doi.org/10.1145/3347318.3355524

1 INTRODUCTION

Analyzing video content to produce summaries and extracting highlights has been challenging for decades. However, the interest for these challenges has lately increased with a focus on sports content. More specifically, soccer is one of the domains that has invested the most in the video analysis field, owing to the massive popularity of the game, to its accordingly huge business market, and lately to the emergence of sport bet companies in many countries.

One of the biggest challenges for automatic soccer video summarization is to produce summaries causing as much emotion as the ones made by human operators. Indeed, a summary cannot be only a sequence of goals, even if goals are important events. Current summaries are produced by professional operators who try to render the story of the match, reflecting the dramaturgy of the match, and possibly connect this story with the last news about the players involved. Producing automatically such kind of summaries is directly impacted by two properties of non-standard soccer video content:

- *the length of a single sample:* to summarize a full match, the most intuitive solution is to consider each match as a single sample. In such a case, the length of the full game (at least 90 minutes, often more) is far from the standard length of samples in benchmark video datasets. Furthermore the ratio between the length of the full game and the target length of the summary (from 20 seconds to 3 minutes, rarely more than 5 minutes) is also very atypical compare to summarization benchmarks.
- *Editorial bias of broadcasted content choice:* State-of-the-art methods to summarize soccer videos and to extract highlights, have been relying on the video broadcasted while this content has already been edited and produced in live by the TV director, and is thus a biased version of the original content of the real game. This bias is for example striking in the diversity of view angles chosen by the director to render two similar actions at different time in the same game. This makes even more difficult the task of recognizing similar actions to further decide which one should be selected to be part of the summary. Videos broadcasted on TV can even miss some events during the replays or the changes of camera.

Faced with these challenges, recent research works have tried to get rid of the producer subjectivity and the amount of data when considering a match as a sample, by benefiting from other data than multimedia content broadcasted on TV. For professional sports with very important economic markets such as soccer, rugby, baseball, ice hockey, basketball, cricket, (american) football, several companies (OPTA, STATS, Instat, Wyscout, Scout7, Catapult, Ortec, ChyronHego, Metrica and many more) have proposed to provide analytics for professional clubs with statistics on the activities (during the games, during the training, focusing on a specific player, evaluating strategies, etc). To do so, these companies rely on precise real time metadata associated to every little atomic *event* by a human operator in the stadium reporting all the events that are indeed occurring on the field. These detailed *event* metadata have recently become of great interest for sport bet gamblers and exploited by companies producing multimedia content for the gamblers and other soccer fans. Thus they are not longer only used to build a precise player or team profile but also to extract highlights from soccer games [5].

This is the reason why we propose a multimodal approach combining the complementary relevance of the audio-visual content on the one hand, and the full event details provided by sports analytics companies like OPTA, STATS or Sportsradar.

It is important to set some terms used on this paper. Usually a piece of video is called *a shot*, but in the context of our work there are also *shots* referring to kick a soccer ball. For this reason, from now on we call *clip* a piece of video and *shot* the action of kicking a ball. We indifferently use the terms game or match. It is also important to clarify the difference between action and event. An *event* is an "atomic" activity such as a pass, a tackle or a shoot while an *action* is a group of events such as a goal opportunity. While the events are atomic objective activities observed on field, actions are usually related with the edition of broadcast companies. A goal opportunity is composed by a group of events but it additionally contains subjective parts of the video like views of the coach, the bench or the crowd reaction.

2 RELATED WORK

Several methods have been proposed for video sports summarization. The most popular approach to detect the highlights in sports videos uses handcrafted features extracted from frames like replay detection, play-break rules, shot cut density, zoom-in and zoom-out detection, face detection, motion analysis, etc [1, 8, 14, 16, 19, 26, 28, 31].

Other methods exploit other type of features: Fiao et al. [10] used the emotions shared by the spectators during the match. In [3, 4, 25], authors proposed automatically generate highlights by analyzing social media data. Other approaches have developed different strategies based on the excitement features, from the players reaction and crowd cheering [17, 22, 23]. Tang et al. [27] used a deep learning architecture to detect and classify the events of a match using the text information from soccer statistics websites.

Although these previous works have demonstrated that multiple features play an important role to summarize or detect the highlights in sport videos, they present some limitations. Most of these methods have based the final decisions on heuristics or sport-specific rules combined with expert system [9, 18].

For video summarization in general, many methods are focused on optimizing the diversity of the resulting summaries. Zhang et al. [33, 34] use the Determinal Point Process, Mahasseni et al. [15] consider an unsupervised generative adversarial network, Zhang et al [35] build a sequence-to-sequence encoder and, in order to not rely on only one objective, the authors of [11, 13] define the summary as a combination of interestingness, uniformity and representativeness. However, the main objective in sports summarization is not the diversity, since usually similar actions are present in the summary. For instance, if the objective is to maximize the diversity of a soccer match summary, probably the algorithm will never choose all the goals, while concatenating all the goals is often considered as a starting point of the most basic solutions.

More recently, Zhao et al. [36] presented a hierarchical LSTM as a solution to capture long-range temporal dependency of videos. They showed that their model can handle up to 1600 frames. However this is far from enough in our context, since a soccer match video contains more than 100,000 frames.

Event metadata from soccer matches has proven to be very relevant to analyze different properties of the game. Some methods have used this data to predict the probability that a particular game state will lead to a goal [5], detected team tactics [6] and predicted the potential of players [29]. Although Decroos et al. [5] also consider event metadata, their main goal is to predict the probability that a given game state will lead to a goal. They assume soccer highlights are represented only by goals or goal attempts. However, yellow card, substitution, etc, could be in the summary. Modifying [5] to solve our problem would mean designing a brand new method.

Our approach is illustrated in Figure 1. It combines precision and relevance of event metadata with the expressiveness of multimedia content. It consists of three stages:

- The *Proposals* stage deals with the similarity of inter-categorical actions. Two very similar sets of events pass, tackle, pass can be parts of two different actions goal-opportunity and corner, the former being in the summary while the latter not. This issue is addressed by a Multiple Instance Learning (MIL) network providing a score for each event, further concatenated to end up with consecutive positive events as proposals.
- The *Summarization* stage consists of a multimodal Hierarchical LSTM. The first level LSTM accumulate in each proposal from previous stage, the emotion and excitement information of every concerned event using metadata-based feature vectors concatenated with audio features. The second level is a bi-LSTM capturing the forward-backward temporal dependencies among proposals (in a storytelling engine fashion) in order to predict the probability of each proposal to be selected into the summary.
- the *Content refinement* stage exploits the visual information of each frame to refine the boundaries of the clips predicted as being part of the summary so that the resulting clips are not anymore restricted to start and/or end on event boundaries. A final bi-LSTM network hence predicts which frames among the ones belonging to the pre-selected proposals should be preserved in the final summary. This last stage allows also to focus on visually salient frames.

3 DEEP MULTIMODAL ARCHITECTURE

A video soccer summary consists of clips which represent the most important actions of the match. Each of these clips is made of several events, e.g. in a goal clip there are other events than the event

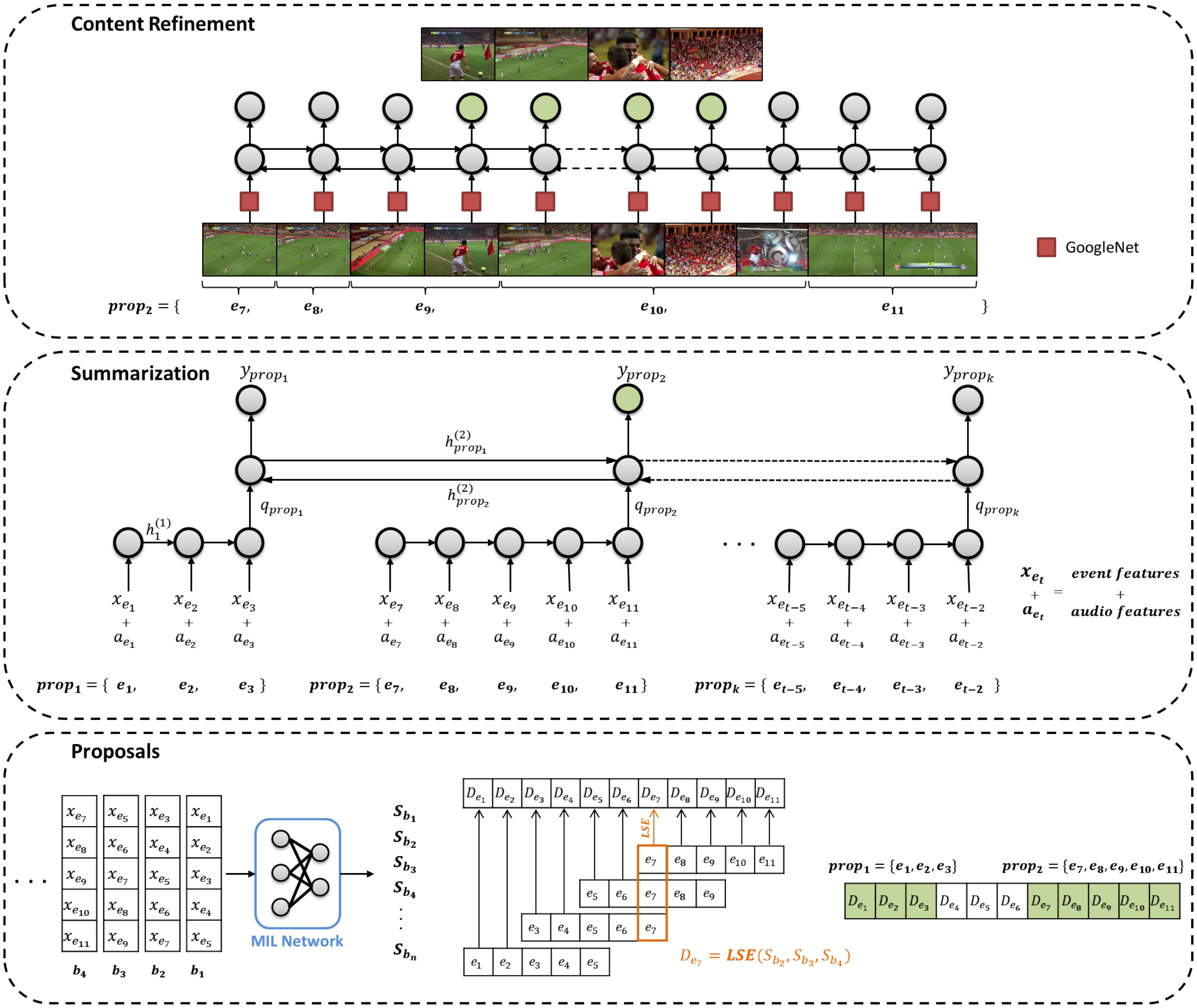

Figure 1: Our proposed approach for soccer video summarization. The Proposals stage takes as input events grouped into bags b_n to be processed by a Multiple Instance Learning (MIL) Network, then Log-Sum-Exp (LSE) function is applied on all the predicted values of each event to finally group the consecutive positive events into Proposals. The Summarization stage is composed by a hierarchical LSTM: bottom LSTM layer creates a representation of each proposal $prop_k$ and an upper bidirectional LSTM decides whether the proposal is part of the summary. The Content Refinement stage takes the frame features of the positive proposals and decide which ones of them indeed belong to the summary.

goal such as the passes, tackles, etc, that led to this goal. The metadata to represent all these events are provided by aforementioned companies like OPTA, Wyscout, STATS, etc.

3.1 Proposals

One of the main challenges for summarizing sport videos is that the same event does not hold the same importance all along the game. Indeed, the same type of event (lets say a pass) is labeled as positive (when leading to a goal) and negative (when leading to nothing) in different parts of the match. Although the goal clip is labeled as positive, it is also composed by some negative events. These characteristics fit with the description of a Multiple Instance Learning (MIL) problem, where a set of instances (in our case events) is labeled as positive if at least one of these instances (one of the events in the set) is positive [7]. In MIL context, the set of instances is called a *Bag*. Negative bags are made of only negative instances (negative events).

We use the MI-Net architecture proposed by Wang et al [30]. This network consists of three fully connected layers followed by one MIL Pooling layer, where the latter aggregates all instance features in order to learn a bag representation. This network receives bags of instances as inputs, and outputs a score per bag.

In this paper, we represent a match as a sequence of events $E = \{e_1, e_2, ..., e_t\}$. These events are "atomic" soccer actions like: pass, tackle, out, interception, post, throw-in, head, etc. If there are three consecutive passes in the game, you will have three similar events "pass" in a row. $X = \{x_{e_1}, x_{e_2}, ..., x_{e_t}\}$ represents the set of instances, where x_{e_t} is the feature vector characterizing the t-th event of the match. We denote by $B = \{b_1, b_2, ..., b_n\}$ the set of bags, where a bag refers to a consecutive subset of instances from X. For a more precise analysis of the match, bags are defined by a sliding window with certain overlap between bags.

We denote the MIL Network output, i.e. the score for the bag b_n, by S_{b_n}. Since an event might belong to several bags owing to the sliding window overlap, we need a method to obtain the accumulated score per event D_{e_t} integrating all the predictions of the bags this event belongs to. Wang et al. [30] evaluated Log-Sum-Exp, Max and Min functions as pooling methods to define the score per bag. We have empirically compared these methods to obtain the score per event and we have found that Log-Sum-Exp function, given in Equation (1), provides the best results.

$$D_{e_t} \geq r^{-1} \cdot log\left[\frac{1}{|\{b_n \mid x_{e_t} \in b_n\}|} \sum_{b_n | x_{e_t} \in b_n} r \cdot S_{b_n}\right] \tag{1}$$

Once we obtain D_{e_t} for each event, we use a threshold T_{ps} to select the positive events. Then we group the consecutive positive events into proposals $P = \{prop_1, prop_2, ..., prop_k\}$, where $prop_k = \{e_t \mid D_{e_t} >= T_{ps}\}$. Thus proposals are most of the time related to what is commonly called actions in sport commentaries.

3.2 Summarization

Multimodal features. Audio plays a very important role in sports, where crowd cheering and excitement in the commentators' tone are usually indicators of an important action. For this reason, our Summarization stage not only use the event metadata features but also the energy of the audio corresponding.

As proposed by Rui et al. [20], we use sub-band short-time energies. Considering the perceptual property of human ears, we can divide into four sub-bands the critical bands that represent cochlear filters in the human auditory model [21]. These sub-bands are 0-630Hz (En_1), 630-1720Hz (En_2), 1720-4400Hz (En_3), and 4400Hz and above (En_4).

For an audio signal A with a sampling rate d, the short-time energy for each sub-band l at any second s is defined as

$$En_l^s = \frac{1}{2*s*d} \sum_{j=s*d}^{3*s*d} A(j) \tag{2}$$

Since each event has a timestamp corresponding to the video time when the event occurs, the energy of the event e_t is the energy of the second s corresponding to its timestamp. We set audio features into a vector a_{e_t} concatenating $En^s, En_1^s, En_2^s, En_3^s$, and En_4^s.

For the Summarization stage, a multimodal instance is then represented by the concatenated feature vector $\{x_{e_t} + a_{e_t}\}$.

Hierarchichal modeling. Hierarchical LSTM has shown to be very efficient for video summarization since it helps to model longer dependencies than traditional LSTM [35–37]. However, previous works use the entire video as input of the first level of the hierarchy but for a lot shorter videos. Instead, because of the size of each video we consider here, we propose to take as input only the relevant parts of the match extracted from our previous Proposals stage.

The Summarization stage is a two-level Hierarchical LSTM. The first level creates a representation of each proposal by accumulating in each proposal the emotion and excitement information of every concerned event; and the second level captures the forward-backward temporal dependencies among proposals (in a storytelling engine fashion) in order to predict the likelihood of each proposal to be selected in the summary.

To be more precise, the input of the first level is $\{x_{e_t} + a_{e_t}\}$, the multimodal feature vector for the $t - th$ event. Assuming e_t belongs to the proposal $prop_k$, the hidden state of this level's LSTM unit is $h_{e_t}^{(1)}$, encoding all events in the proposal $prop_k$ up to the $t - th$ event by computing over the current feature vector $\{x_{e_t} + a_{e_t}\}$ and the previous hidden state $h_{e_{t-1}}^{(1)}$.

After processing an entire proposal, we denote the final hidden state of the LSTM unit as q_{prop_k}, the encoding vector for the proposal $prop_k$. The LSTM unit memory and the initial hidden state are then reset to zero.

After all the proposals are processed, we end up with a sequence of encodings $Q = \{q_{prop_1}, q_{prop_2}, ..., q_{prop_k}\}$. We construct a bidirectional-LSTM over Q which grasps the temporal dependencies between proposals in the summary.

Indeed, in a game with a largely unbalanced score (8 to 1 for instance) the summary may not present all the goals; or in a game with not much action, with many shots-not-on-target, the summary may contain only some of them maybe the first ones or maybe some evenly distributed with respect to other actions present in the summary.

These choices are directly related to the story-line of the summary that we target to learn with this bidirectional-LSTM.

The output of this second level is denoted as $Y = \{y_{prop_1}, y_{prop_2}, ..., y_{prop_k}\}$, where y_{prop_k} indicates the likelihood of whether the proposal $prop_k$ should be included in the summary.

We use a threshold T_{ss} to select which are the proposals that comprise the predicted summary. We denote this summary by $Summ = \{prop_k \mid y_{prop_k} >= T_{ss}\}$

3.3 Content Refinement

Representing a match as a sequence of events significantly reduces the amount of information to process compared to analyze the content at frame level. However as mentioned before, the events are occurring on the field since they are acquired by a person watching the game in the stadium and not behind a TV screen. Hence, most of the times the boundaries of the clips (on TV) are not completely aligned with on-field events, since clips boundaries are decided by the producer who might cut the content in the middle of an event. Thus, the final clips cannot be restricted to start and/or end on event boundaries.

For this reason, we decided to exploit the visual content broadcast on TV by training a final bidirectional-LSTM to decide which frames of the selected events through *Summ* really belong to the final summary. This last stage introduces visual features in the process and thus accounts for new features to capture possible interestingness and representativeness within the proposal.

Let's define $G = \{g_1, g_2, ..., g_n\}$, where each g_n is a positive proposal extracted from the Summarization stage. The beginning and the end of g_n is given by the timestamp of the first and last events of g_n. We represent the frame corresponding to this beginning and end events by $F_{g_n}^{beg}$ and $F_{g_n}^{end}$ respectively.

Each input sample of the LSTM corresponds to the feature vectors of the frames inside the interval $[F_{g_n}^{beg}, F_{g_n}^{end}]$. And for each of these frames there is an output, which corresponds to the likelihood of this frame to be part of the final summary.

4 EXPERIMENTS

We first introduce the experimental setting, describing the dataset, features and metrics. We then present the quantitative results to demonstrate the advantages of the proposed approach over a frames-based method and the comparison of our model with and without audio features. We further perform a qualitative comparison.

4.1 Setup

Summary-based Dataset. Our dataset consists of 20 complete soccer games from 2017-2018 season of French Ligue 1. The ground truth video summaries were made by professional editors of a sports broadcast company. The matches were played at different times of the day, in multiple fields and with 19 different teams. We have manually detected the corresponding temporal intervals of each summary clip in the corresponding original matches. We create 5 folds where each fold has 16 games for training and 4 games for testing.

In order to create our event-based dataset, we use event metadata provided in real time by sport analytics companies. Usually these companies provide a time coded feed that lists the action events within a game each associated with a player, a team, an event type, an event qualifier, outcome, x and y position on the field, minutes and seconds. Some examples for event type are: pass, tackle, foul, out, corner, interception, save, miss, post... Some qualifiers of the event *pass* are long ball, free-kick, throw-in, assist, kick-off. Outcome has value 1 if the action was satisfactory finished (after a tackle, the player dispossesses an opponent of the ball and retain possession) or 0 otherwise.

The first step to create the dataset for the Proposal stage is to find the candidates. We take all the summary clips and extract the sequences of events corresponding to these clips. For instance, in one of the summaries created by the editors, there is a goal clip which corresponding pattern is: out, throw-in, long ball, aerial, pass, goal. Then we look for this exact same sequence in the rest of the match and in the all remaining matches in the training set to annotate them as ground truth proposals. The second step is to create the bags, we slide a window of 5-events size and stride 2. The bag is considered as positive if at least 3 of the 5 events belong to a ground truth proposal. It is important to notice that to find the ground truth proposals of the training set, we only look for event sequence patterns present inside this set (not in the test set).

For the Summarization stage, a proposal $prop_k$ is considered as part of the summary if at least one event of the proposal is overlapped with a clip of the ground truth video summary.

Action-based Dataset. We have created a soccer action dataset, which is different from the previous Summary-based dataset. The temporal annotations were made in real-time by professional editors i.e. while the matches were broadcasted. Furthermore, the annotations are defined at action level, that means at a larger scale than event level considered before, leading to 20 different actions: *Start Match, End Match, End First Half, Start Second Half, Saved Field, Corner, Shot on Target, Shot not Target, Penalty Missed, Goal on Penalty, Goal on Field, Yellow Card, Free-kick, Substitution, Post, Offside, Red Card, Yellow Card, Chance, Injury.* In our 20 games, these actions are in average 30 seconds long but they can vary from 7 to 85 seconds.

It is important to note that the standard benchmark datasets for action detection differ in several aspects with this soccer dataset. Event though these benchmarks contain samples with multiple actions inside, it is very common to find the case where all the actions in a video sample belongs to the same class. In ActivityNet [2], only 0.06% of the video samples contain at least two different classes of actions and for THUMOS-14 [12] this ratio is 15%. While in our soccer dataset, a sample is a video match which always contains many type of actions. In addition, a significant part of these video match samples is Background (parts of the match where nothing important occurs). The actions represent only 22% of our dataset, compared with 31% in THUMOS-14 and 49% in ActivityNet.

Features. The feature vector x_{e_t} is the concatenation of: one hot-encoding vector for the type of event, one hot-encoding vector for the event qualifier, outcome value, x position, y position, and the time passed from the previous event in seconds.

As mentioned before, the feature vector a_{e_t} is the representation of the audio signal energy. We concatenate the energy for all the frequencies followed by the energy of each sub-band.

For the Content Refinement stage each frame is represented by the output of the penultimate layer (pool 5) of GoogleNet [24] (1024-dimension feature vector).

Networks specifications. The Multiple Instance Learning network has 256, 128 and 64 neurons in its three fully connected layers. Each of the two LSTM layers of the Summarization stage have 128 units. The LSTM of the Content Refinement stage has 256 units. To train our model, we adopt a stage-wise routine, using Adam optimizer and binary cross-entropy as loss function.

Evaluation. As in previous works on video summarization [15, 33, 35], we evaluate our generated summary U computing the Precision, Recall and F-score against V, the summary created by the editors:

$$\begin{aligned} Precision &= \frac{\text{overlapped duration of } U \text{ and } V}{\text{duration of } U} \\ Recall &= \frac{\text{overlapped duration of } U \text{ and } V}{\text{duration of } V} \\ F_{score} &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \end{aligned} \tag{3}$$

4.2 State-of-the-art Summarization Methods

One challenge in our context lies in comparing our approach to the state-of-the-art, relevant datasets are under copyright infringement

Method	Precision	Recall	F-score
event-vsLSTM (with audio)	0.414	0.389	0.384
event-H-RNN (with audio)	0.257	0.594	0.355
Ours (with audio)	0.470	0.457	0.459

Table 1: Multimodal Performance Comparison. All the models were trained with event and audio features.

and as far as we know no other method considers a full soccer game as a single sample.

As mentioned in Section 2, other methods are not suitable for comparison. Methods [11, 13, 15, 33–35] optimize summary diversity which is not convenient for soccer videos since a summary could contain several similar actions; methods in [3, 4, 10, 17, 22, 25, 27] use different input multimedia data (text or comments from social networks) not easily reachable.

vsLSTM [34] and H-RNN [36] appeared to be the only summarization methods we have found where input samples are full videos, optimization does not rely on diversity and input data is similar to ours. Therefore, we trained these two methods from scratch using the frame features extracted on our video dataset.

H-RNN. This approach splits the video on fixed-size segments and use as input frame features extracted from GoogleNet. We set the segment size to 40, as recommended by the authors.

vsLSTM. The model is a bidirectional-LSTM followed by a multilayer perceptron. The inputs are frame features extracted from GoogleNet.

In addition, we compared with two models which take the same ideas of H-RNN and vsLSTM but modified to be compliant with the inputs of our approach.

event-H-RNN. As the original H-RNN we use fixed-size segments but instead of using frame features as inputs, this new model takes the same input as our approach, either x_{e_t} or $\{x_{e_t} + a_{e_t}\}$ (in that last case we precise "with audio" in the experiments). Although the authors of H-RNN advise to use 40 as segment size, we have empirically found that for our event-based approach is better to choose a significantly smaller size, most likely due to the fact the mean number of events on a ground truth clip is 7. For this reason and to perform a fair comparison, we have used as segment size the bag size used in our Proposals stage.

event-vsLSTM. It is a vsLSTM architecture but instead of using frames features, it takes the same input as our approach, either x_{e_t} or $\{x_{e_t} + a_{e_t}\}$ (in that last case we precise "with audio" in the experiments).

4.3 Performance Results

All the scores reported in this section correspond to the results of the 20 games of our dataset, we gather the test sets results of the 5 folds.

Comparison in a multimodal context. In Table 1, we compare our approach with event-H-RNN and event-vsLSTM, since they are our multimodal models. To obtain these results, the event models were trained with the concatenation of the events and audio features $\{x_{e_t} + a_{e_t}\}$ as input, and our approach was trained as explained in the previous section. The input for the Proposals stage is the event feature x_{e_t}, for the Summarization stage it is $\{x_{e_t} + a_{e_t}\}$ and for the Content Refinement stage it is the frame feature. Event-H-RNN shows the highest Recall but the Precision is the lowest, which means this method has issues to identify the events that do not belong to the summary. Our approach clearly outperforms in terms of F-score and Precision, it also shows a good trade-off between Precision and Recall.

The results of the state of the art show that H-RNN performs better than vsLSTM [35–37], however with our data event-vsLSTM obtains better F-score and Precision than event-H-RNN. Probably the fixed size of the segment and the overlap to decide if the segment is positive, have to be carefully analyzed. We have tried different segment sizes and overlap ratio, and report the best results.

Method	Precision	Recall	F-score
vsLSTM	0.553	0.241	0.296
H-RNN	0.406	0.295	0.335
Ours	0.436	0.401	0.415

Table 2: Performance comparison with frames based models. H-RNN [36] and vsLSTM [34] were trained with frames features, as it was proposed originally in the papers

Comparison with frame-based models. To verify that our approach is better than frame-based methods, the results of H-RNN and vsLSTM are provided in Table 2. To be fair, we make the comparison with our approach that was trained without audio features. Although vsLSTM has higher Precision, its Recall is the lowest. One possible interpretation for this case (that we also checked visually on the resulting summaries) is that the method only learned to correctly predict the most common actions like shots on target. The Recall and F-score of our method are the highest, and it is the method with the best trade-off between Precision and Recall. This clearly shows that even without the audio features our event-based method is able to extract the most accurate summaries.

Focus on missing clip and false negative rates. Since the last stage of our approach is only in charge of the refinement of clips predicted as summary, it is very important that the Summarization stage misses the least number of clips belonging to the summary. The column *Missing clips* of Table 3 represents the ratio between the number of clips that where not detected at all and the total number of clips in the summary. Our method has the lowest missing clips ratio, with almost 10% less than the second best on this column. Table 3 also reports the Recall and false negatives, where the latter is the ratio between the seconds that were not detected and the total duration of the ground truth summary. Our approach gets the lowest ratio of false negatives and highest recall. All these results prove that our combination of Multiple Instance Learning and hierarchical LSTM is the best choice because no matter how good our Content Refinement stage is, if we replace our first two

Method	Missing Clips	False Negatives	Recall
vsLSTM	0.509	0.759	0.241
H-RNN	0.548	0.705	0.295
event-vsLSTM (with audio)	0.364	0.611	0.389
event-H-RNN (with audio)	0.398	0.406	0.594
Summarization stage (with audio)	0.267	0.355	0.644

Table 3: Comparison on undetected parts of ground truth summary. Missing clips represent the percentage of clips which where completely missed. And False Negatives represent the percentage of all seconds that were not detected.

Method	Precision	Recall	F-score
event-vsLSTM	0.435	0.351	0.381
event-H-RNN	0.249	0.567	0.337
Ours	0.436	0.401	0.415
event-vsLSTM (with audio)	0.414	0.389	0.384
event-H-RNN (with audio)	0.257	0.594	0.355
Ours (with audio)	0.470	0.457	0.459

Table 4: Performance comparison for models with and without audio features.

stages by any of the state-of-the-art models, these models will always provide less and possibly shorter positive proposals.

Impact of audio features. It is worth mentioning that audio features play an essential role to detect important actions. If we compare the results between the models trained with audio and the ones trained without audio (shown in Table 4), we can see that adding the audio features usually improves the scores, especially the Recall. In addition, our approach performs the best in terms of F-score, even without these additional features.

4.4 Summarization as action classification

One could think about the soccer game summarization as a general video classification task. Such an approach seems very reasonable considering the late performances of video classification methods such as R-C3D [32]. Since R-C3D aims at detecting actions in videos, this approach could be a very good candidate to replace our Proposals stage. For this reason we have trained R-C3D to detect all the actions of a match.

We have thus compared the percentage of missing proposals of our approach with the percentage of missing actions of R-C3D. As described in section 4.1 (Summary-based Dataset), to extract the ground truth proposals for the Proposals stage dataset we take the sequences of events labeled as summary and look for the same sequences on the entire set. The dataset used for R-C3D, detailed in section 4.1 (Action-based Dataset), is composed by several soccer actions labeled by professional editors.

To evaluate R-C3D on our soccer action dataset, we used the code and weights published by the authors. We chose to initialize the network with the weights trained on THUMOS-14 because among the datasets they have explored, THUMOS-14 is the dataset with the closest properties to ours (i.e. ratio actions-vs-Background). We have randomly chosen 80% of our 20 matches for training and the remaining 20% for testing.

We have found that R-C3D misses 84 actions from the 266 total test actions of our soccer action dataset, which represents 32%. While our Proposals stage misses only 7.98%, 34 out of 426 proposals of the test set.

4.5 Qualitative Results

We illustrate an example of our video summarization results in Figure 2 in order to show how important is the Proposals and Summarization stage to obtain good results in our Content Refinement stage. The rectangles below the images represent the time-line across the match, the inner colored rectangles are time intervals and the images above the rectangles are frames sampled from the match video. The top picture with the green time line shows two different examples of ground truth summary intervals. The middle and bottom picture depict the Summarization and Content Refinenment predictions that are closest to the ground truth intervals.

The example on the left side shows a good behavior of our model, where the Proposals stage detects all the events surrounding a ground truth interval, the Summarization stage classifies as positive this proposal and finally the Content Refinement stage is able to detect the most relevant part of the clip. The example on the right side of the figure presents a case where the first stage of our method creates a proposal that misses the beginning of the ground truth interval, however the last stage is able to predict that the beginning of the proposal is relevant for the summary.

The sampled frames of Figure 2 are also important to show that the borders of the ground truth intervals might be subjective. On the left-side example, the beginning of the clip is when the player kicks the ball from the corner and the end is the coach reaction, our prediction starts before when the camera shows a wide angle of the corner shot and finishes when the player approaches for a throw-in. One could argue that both the predicted and the true borders are valid. A similar situation occurs at the end of the right-side example, the ground truth clip ends on the team celebration and the prediction ends the clip on the audience celebration.

Figure 3 depicts the prediction intervals for an entire match, where the top picture with the green rectangle represents the ground truth intervals. As shown on this second row (from top to bottom) the Proposals stage is able to detect all the intervals from the real summary. And the third row demonstrates that although the Proposals stage outputs multiple false positives, the

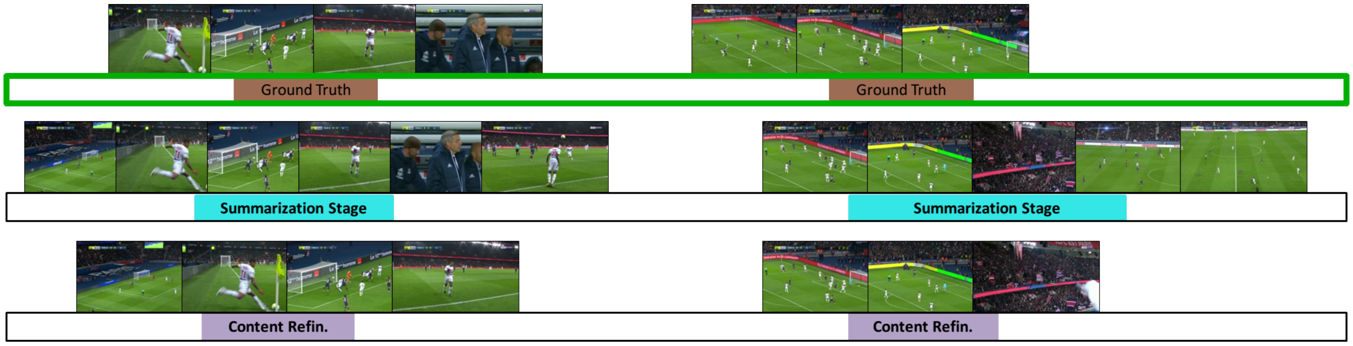

Figure 2: Video prediction example of our method. Pictures on the top are sampled from the ground truth summary, ones in the middle are from the Summarization stage and in the bottom are from the final summary prediction of our method. The color bars below the images represent time intervals.

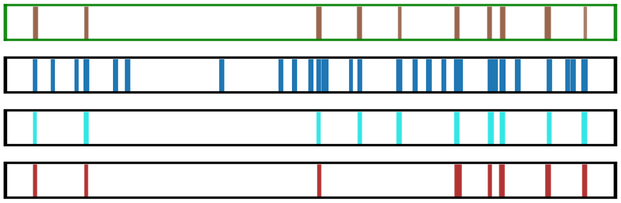

Figure 3: Comparison of intervals prediction of one entire match. The topmost row shows the ground truth intervals. Results of the Proposals and Summarization stage are the second and third rows respectively. The bottom row shows the intervals prediction of event-H-RNN model.

Summarization stage is able to detect which proposals indeed belong to the summary. The bottom picture represents the prediction intervals of event-H-RNN model, where we can visually confirm the high rate of missing clips and false negatives, shown in previous section.

Another important property of our model that is worth emphasizing is the fact that the Multiple Instance Learning Approach significantly helps to extract meaningful proposals. There is a particular interval that corresponds to a substitution, which is a very uncommon action in our summaries. Our Proposals stage is able to capture this event even when there is no substitution interval in the training set.

5 DISCUSSION

Our method requires to set only two parameters T_{ps} and T_{ss}, which are actually not very sensitive, usually the scores of the positive samples are clearly higher than the scores of the negatives ones. Furthermore, our approach can be seen as a computer-aided summarization platform which will help the current human operators to build summaries.

Our method is not restricted to soccer since all the aforementioned companies providing precise event metadata are already providing similar information for several sports with large audiences (basketball, ice hockey, tennis, rugby, american football...). We will investigate in the future how easily our method can be transferred to other sports and at what cost.

The previous question remains at amateur levels where we cannot get access to similar rich metadata. In that context, our intuition is that considering events as the right semantic level of content analysis looks better compared to actions. We could design a new stage, located before our Proposals stage, that uses frames to find the different events of the match. The rest of the architecture could remain the same.

6 CONCLUSION

We presented a novel approach for automatic generation of summaries from soccer videos based on multimodal features, including audio energy, event features from sports analytics and visual information from video frames. The proposed approach consists of three consecutive stages: a Proposals stage deals with the similarity between the events inside the summary and the rest of the match, a Summarization stage accumulates the emotion and excitement information of each proposal to capture the temporal dependencies in order to decide which proposals are part of the summary and, finally a Content Refinement stage exploits the visual information to predict which frames among the ones belonging to the pre-selected proposals should be preserved in the final summary. Our model outperforms by an 8% margin not only the video processing state-of-the-art methods but also methods that use event and audio features. There are several key contribution factors: the capacity of Multiple Instance Learning to deal with similar inter-categorical actions, our idea to complement hierarchical LSTMs' strength with the generation of proposals and, the use of audio features to improve the detection of important events. While other methods propose to use frames, we have demonstrated that to base the summarization on events get a shorter and better representation of longer videos as soccer matches.

7 ACKNOWLEDGMENTS

This work has been co-funded by Région Provence Alpes Côte d'Azur (PACA), Université Côte d'Azur (UCA) and Wildmoka Company.

REFERENCES

[1] Vinay Bettadapura, Caroline Pantofaru, and Irfan Essa. 2016. Leveraging contextual cues for generating basketball highlights. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 908–917.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[3] Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[4] David Corney, Carlos Martin, and Ayse Göker. 2014. Two Sides to Every Story: Subjective Event Summarization of Sports Events using Twitter.. In *SoMuS@ ICMR*. Citeseer.

[5] Tom Decroos, Vladimir Dzyuba, Jan Van Haaren, and Jesse Davis. 2017. Predicting soccer highlights from spatio-temporal match event streams. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[6] Tom Decroos, Jan Van Haaren, and Jesse Davis. 2018. Automatic Discovery of Tactics in Spatio-Temporal Soccer Match Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 223–232.

[7] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89, 1-2 (1997), 31–71.

[8] Ahmet Ekin, A Murat Tekalp, and Rajiv Mehrotra. 2003. Automatic soccer video analysis and summarization. *IEEE Transactions on Image processing* 12, 7 (2003), 796–807.

[9] Mohamed Y Eldib, Bassam S Abou Zaid, Hossam M Zawbaa, Mohamed El-Zahar, and Motaz El-Saban. 2009. Soccer video summarization using enhanced logo detection. In *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 4345–4348.

[10] Guilherme Fião, Teresa Romão, Nuno Correia, Pedro Centieiro, and A Eduardo Dias. 2016. Automatic Generation of Sport Video Highlights Based on Fan's Emotions and Content. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*. ACM, 29.

[11] Michael Gygli, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3090–3098.

[12] Yu Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. 2014. THUMOS challenge: Action recognition with a large number of classes.

[13] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. 2017. A general framework for edited video and raw video summarization. *IEEE Transactions on Image Processing* 26, 8 (2017), 3652–3664.

[14] Tingxi Liu, Yao Lu, Xiaoyu Lei, Lijing Zhang, Haoyu Wang, Wei Huang, and Zijian Wang. 2017. Soccer video event detection using 3D convolutional networks and shot boundary detection via deep feature distance. In *International Conference on Neural Information Processing*. Springer, 440–449.

[15] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 202–211.

[16] Engin Mendi, Hélio B Clemente, and Coskun Bayrak. 2013. Sports video summarization based on motion analysis. *Computers & Electrical Engineering* 39, 3 (2013), 790–796.

[17] Michele Merler, Dhiraj Joshi, Quoc-Bao Nguyen, Stephen Hammer, John Kent, John R Smith, and Rogerio S Feris. 2017. Automatic curation of golf highlights using multimodal excitement features. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 57–65.

[18] Ngoc Nguyen and Atsuo Yoshitaka. 2014. Soccer video summarization based on cinematography and motion analysis. In *2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.

[19] Arnau Raventos, Raul Quijada, Luis Torres, and Francesc Tarrés. 2015. Automatic summarization of soccer highlights using audio-visual descriptors. *SpringerPlus* 4, 1 (2015), 301.

[20] Yong Rui, Anoop Gupta, and Alex Acero. 2000. Automatically extracting highlights for TV baseball programs. In *Proceedings of the eighth ACM international conference on Multimedia*. ACM, 105–115.

[21] Bertram Scharf. 1970. Critical bands. *Foundation of modern auditory theory* 1 (1970), 159–202.

[22] Pushkar Shukla, Hemant Sadana, Apaar Bansal, Deepak Verma, Carlos Elmadjian, Balasubramanian Raman, and Matthew Turk. 2018. Automatic Cricket Highlight generation using Event-Driven and Excitement-Based features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1800–1808.

[23] Mohamad-Hoseyn Sigari, Hamid Soltanian-Zadeh, and Hamid-Reza Pourreza. 2015. Fast highlight detection and scoring for broadcast soccer video summarization using on-demand feature extraction and fuzzy inference. *International Journal of Computer Graphics* 6, 1 (2015), 13–36.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[25] Anthony Tang and Sebastian Boring. 2012. # EpicPlay: Crowd-sourcing sports video highlights. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1569–1572.

[26] Hao Tang, Vivek Kwatra, Mehmet Emre Sargin, and Ullas Gargi. 2011. Detecting highlights in sports videos: Cricket as a test case. In *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.

[27] Kaiyu Tang, Yixin Bao, Zhijian Zhao, Liang Zhu, Yining Lin, and Yao Peng. 2018. AutoHighlight: Automatic Highlights Detection and Segmentation in Soccer Matches. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 4619–4624.

[28] Mostafa Tavassolipour, Mahmood Karimian, and Shohreh Kasaei. 2014. Event detection and summarization in soccer videos using bayesian network and copula. *IEEE Transactions on circuits and systems for video technology* 24, 2 (2014), 291–304.

[29] Ruben Vroonen, Tom Decroos, Jan Van Haaren, and Jesse Davis. 2017. Predicting the Potential of Professional Soccer Players.. In *MLSA@ PKDD/ECML*. 1–10.

[30] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. 2018. Revisiting multiple instance neural networks. *Pattern Recognition* 74 (2018), 15–24.

[31] Zengkai Wang, Junqing Yu, Yunfeng He, and Tao Guan. 2014. Affection arousal based highlight extraction for soccer video. *Multimedia Tools and Applications* 73, 1 (2014), 519–546.

[32] Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-C3D: region convolutional 3d network for temporal activity detection. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 5794–5803.

[33] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1059–1067.

[34] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European conference on computer vision*. Springer, 766–782.

[35] Ke Zhang, Kristen Grauman, and Fei Sha. 2018. Retrospective Encoders for Video Summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 383–399.

[36] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2017. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 863–871.

[37] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2018. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7405–7414.