2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Hierarchical Multimodal Attention for Deep Video Summarization

Melissa Sanabria
Inria, CNRS, I3S, Maasai
Université Côte d'Azur
Sophia-Antipolis, France
sanabria@unice.fr

Frédéric Precioso
Inria, CNRS, I3S, Maasai
Université Côte d'Azur
Sophia-Antipolis, France
precioso@unice.fr

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

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9413097

***Abstract*—The way people consume sports on TV has drastically evolved in the last years, particularly under the combined effects of the legalization of sport betting and the huge increase of sport analytics. Several companies are nowadays sending observers in the stadiums to collect live data of all the events happening on the field during the match. Those data contain meaningful information providing a very detailed description of all the actions occurring during the match to feed the coaches and staff, the fans, the viewers, and the gamblers. Exploiting all these data, sport broadcasters want to generate extra content such as match highlights, match summaries, players and teams analytics, etc., to appeal subscribers. This paper explores the problem of summarizing professional soccer matches as automatically as possible using both the aforementioned event-stream data collected from the field and the content broadcasted on TV. We have designed an architecture, introducing first (1) a Multiple Instance Learning method that takes into account the sequential dependency among events and then (2) a hierarchical multimodal attention layer that grasps the importance of each event in an action. We evaluate our approach on matches from two professional European soccer leagues, showing its capability to identify the best actions for automatic summarization by comparing with real summaries made by human operators.**

***Index Terms*—Event stream data, Soccer match data, Video Summarization, Multimodal data, Sports Analytics**

I. Introduction

The consumption of multimedia content has drastically evolved in the last decade. It is well-known that the amount of multimedia content stored, produced, published, exchanged on internet, is continuously increasing for more than ten years now, with a particular focus on videos. However the amount is not the only parameter which increases, more and more multimedia content evolves towards being more user immersive. This evolution of multimedia content consumption is also noticeable in the field of sport broadcasting where content augmentation improves immersive user experience, providing the user with all possible modality viewpoints of the content so as to enhance user engagement in the media. For instance, in the main world cycling races, sensors provide bike speed, pedaling frequency, heart rate, power produced, by cyclists, allowing the viewers to focus on specific content targeting a specific competitor. It is the same with soccer where new analytics about displacements on the field, run speed, run distance, heart rate, provide the viewer with an augmented experience of the matches.

In addition, specifically in the sport domain, the legalization and democratization of sport betting companies (recently in Europe) have accentuated the need for augmenting even more live content on players, teams, games, etc. Sport broadcasters aim at appealing and retaining subscribers to watch always more sports on TV, providing them with an enriched user experience. All the main sport broadcasters mandate companies, like Opta, Wyscout, Instat, Sportradar, Gamebreaker, SportsCode and many more, to send human observers in the stadiums to collect live data of all the events happening on the field during the match (pass, shoot, foul, players' position, cards, substitutions...) . Those event data contain meaningful information providing a very detailed description of all the actions occurring during the match to "feed" coaches and staffs, fans, spectators, viewers, and gamblers. Sport broadcasters want also to generate extra multimedia content such as match highlights, match summaries, etc, for their subscribers. There is no fully automatic solution to produce such content augmentation, in a short time (targeting real-time), on real sport content (one soccer match to be summarized is at the very least a 90 minute long video). The current solution for sport broadcasters is thus to rely on human operators to generate in live, highlights, summaries, specific content for social networks, and any extra content that will build viewer loyalty.

In this paper, we want to exploit the event data which provide us with unique information compared to all other available modalities (audio, visual, text...) to build a deep architecture for automatic summarization. Furthermore, event data are lighter in memory than video modality and thus allow us to process faster each whole match.

Contributions:

- *Proposal Generation:* Event metadata are first exploited to detect relevant categories of actions (i.e. segments of matches). Based on the intra-category diversity and inter-category possible similarity, we approach this first step as a Multiple Instance Learning problem (MIL) and address it through a LSTM MIL pooling schema.

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

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

• *Hierarchical multimodal attention:* Actions proposed to be in the summary, are classified into belonging or not to the summary. We design for this second step a specific multimodal attention model which combines event and audio modalities, at the event level, in a hierarchical LSTM network.

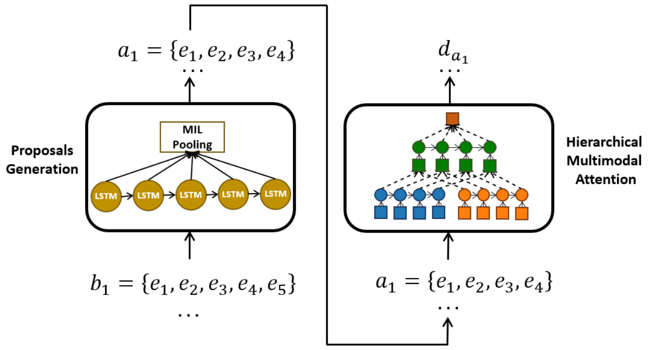

$a_1 = \{e_1, e_2, e_3, e_4\}$
...
Proposals Generation
MIL Pooling
LSTM
LSTM
LSTM
LSTM
LSTM
$b_1 = \{e_1, e_2, e_3, e_4, e_5\}$
...
d_{a_1}
...
Hierarchical Multimodal Attention
$a_1 = \{e_1, e_2, e_3, e_4\}$
...

Fig. 1: General schema of our approach. The left part of the figure represents the first block of our approach: Proposals Generation with a LSTM MIL Pooling. It gets as input the bags of events and outputs action Proposals. The right part of the figure is the second block of our approach: Hierarchical Multimodal Attention. It gets as input the action proposals (events data and audio data) and predict the likelihood for the given action to be in the summary.

II. Related Work

Event data. While our approach uses metadata acquired in live during the matches for summarization, several approaches use this same information for other tasks like recognizing teams [1], analyzing advantage of playing on your home field [2], automatically discovering patterns in offensive strategies [3], [4], predicting passes [5], detecting tactics [6], predicting the chance to score the next goal [7] and evaluating the performance or contributions of the players [8]–[10].

In addition, these metadata can now more and more be found either on program websites directly managed by the companies producing them (Prozone, GeniusSports, Opta, WyScout, and others) or through other open data sources (Kaggle competition, open datasets, etc) [11], [12].

Sports Video Summarization. Early works in video summarization for sports mainly relies on hand-crafted heuristics. They exploit the characteristics of the field (lines, goal mouth), cinematographic properties like the camera motions, slow motion or zooming, and also specific edition patterns like the replays [13]–[16] to select representative frames. More recently, approaches have migrated to machine learning techniques. Liu et al. [17] use 3D convolutional networks to classify the different clips of soccer videos, and Agyeman et al. [18] use this type of networks as feature extractor to then train an LSTM for action classification. Javed et al. [19] propose an extreme learning machine to detect key-events based on the replay information. The main limitations in the state-of-the-art of video sport summarization is the lack of standardization in the evaluation process and the use of heuristics to make decisions. Many of the aforementioned papers do not evaluate their methods using summarization metrics, they usually rather focus on the detection of most important actions like goals.

Multimodal Sports Summarization. For sports summarization the video is not the only source of information. Some methods propose to use social networks like the tweet streams during the game [20]–[22]. Mend et al. [23] summarize a match detecting the intervals with highest motion from the optical flow. Tang et al. [24] design a deep learning algorithm to classify soccer actions from the text timeline provided by several web pages. Other methods exploit audio features [25], [26] since they help to identify the excitement of the commentators and the crowd, and sometimes the ball hit like for tennis or baseball. Multiple modalities play an important role to choose the best moments of sports videos. Several methods [27]–[30] merge different modalities like the sound energy, the score, camera motions, players' reactions, referee whistle, etc. In this work, we are also going to consider different modalities but merging them hierarchically.

Video Summarization. Most of the works on video summarization are not specifically dedicated to sports domain and its peculiarities. For more general purpose videos, approaches focus on different criteria. For instance, the observation that similar videos share similar summary structures [31]–[33]. Taking as inspiration semantic segmentation, Rochan et al. [34] use a fully convolutional network across time, where the output is a mask showing the relevant frames for the video summary. On the other hand [35], [36] use a combination of objectives like interestingness, uniformity, representativeness to identify the most appealing moments. Recent successes of Generative Adversarial Networks have led to several works based on unsupervised approaches for video summarization [37], [38]. Zhang et al. [31] were the first ones using LSTM for video summarization, their method is a bidirectional LSTM followed by a Multi-Layer Perceptron. Although LSTM is able to model long-range structural dependencies, Zhao et al. [39] propose a hierarchical LSTM to help the model to handle particularly long sequences. Most of the methods on the summarization of general purpose videos are based on the maximization of diversity, trying to minimize the number of similar shots [37], [40], [41]. However, such a criterion does not apply for sport summarization. For instance in soccer, the easiest way to create a summary is to choose the goal clips, even if they are all visually very similar. And this situation holds for many sports.

To the extent of our knowledge, our method is the first one exploiting event data to automatically generate summaries. Event data significantly reduces the amount of information to be processed per match compared with other modalities like video where the number of frames in a 90 minutes match makes video data intractable. We propose a Multiple Instance Learning approach that, unlike comparable approaches, exploits LSTM sequentiality to process time-dependent instances and generate proposals. Furthermore, this work also introduces a new multimodal attention architecture that helps to learn a multimodal representation at the event level instead of learning a representation at the action level as existing methods.

III. Proposal Generation

The goal of the first block of our method is to identify the action proposals of the match, that is to say consecutive relevant events. For instance, an action of a *goal* might be corresponding to the sequence $\{$*pass, interception, pass, goal*$\}$. Such groups of events are considered as proposals if they are parts of the match that might belong to the summary.

In the last few years, progress has been made in the task of object detection and one of the common core elements of all these approaches is to split the process in two tasks, object proposal generation as a preprocessing stage that later provides candidate windows to an object classifier. Examples of this are the architectures based on Region Proposal Network (RPN) [42]–[44]. This concept was later developed for videos [45]–[47]. In this paper we follow the same idea of splitting the detection in two tasks, first the generation of proposals and second the classification of these proposals.

However, we believe that in our context, a traditional learning method is not enough. In the same match we can find identical sequences of events, some labeled as positive (i.e. in the summary) and some as negative (i.e. not in the summary). In the case of soccer for instance, the sequence *pass, interception, pass* can be the beginning of an action of *goal* which is part of the summary but the same sequence can belong to some section of the match where nothing relevant is happening.

We thus decide to tackle the similarity of inter-categorical actions with a Multiple Instance Learning (MIL) approach.

As it will be detailed further in Section V-A, our ground truth dataset contains only the parts of matches belonging to summaries which means that we do not have access to the labels of the actions. To tackle this issue, we define as *candidate* all the sequences of events that are identical to the ones labeled as summary. To be more specific, for instance in one of the summaries of the ground truth, there is a goal action which corresponds to the sequence of events = $\{$*out, throw-in, long ball, aerial, pass, goal*$\}$. We then look for all the instances of this exact same sequence of events in the rest of the match and in all the remaining matches in the training set to label them as *candidates*.

We follow the event and bag representation proposed in [48]. A match is a sequence of events $\{e_1, e_2, ..., e_N\}$ which are all the events occurring on the field (possibly not broadcasted on TV). $X = \{x_{e_1}, x_{e_2}, ..., x_{e_N}\}$ represents the set of instances, where x_{e_n} is the feature vector characterizing the n-th event of the match. We denote a bag b as a set of consecutive events and B as the set of bags in a match $B = \{b_1, b_2, ..., b_F\}$. A bag is considered positive if at least half of the events of the bag belongs to any of the *candidates*.

As in test phase we do not have access to the ground truth intervals, the bags are created in a class-agnostic way, using a sliding window with overlap across the entire match.

In the classical supervised learning problem the objective is to find a model that predicts a target value $y \in \{0, 1\}$, for a given instance. In the case of MIL, instead of a single instance there are groups of instances called *bags*. There is also a single binary label Y associated with the bag. Furthermore, it assumes that individual labels exist for the instances within a bag, i.e., $y_1, ..., y_k$ and $y_k \in \{0, 1\}$, however there is no access to those labels and they remain unknown during training. Then in MIL, a bag is labeled as negative if all the instances inside the bag are negative and a bag is labeled as positive if at least one instance of the bag is positive:

$$Y = \begin{cases} 0, & \text{iff } \sum_k y_k = 0, \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

A. LSTM MIL Pooling

MIL is classically regarded as a general and abstract learning paradigm and, as such, it does not require or involve any feature extraction process. However, recent works integrate the process of solving MIL problem, within the process of learning features by using a fully-connected neural network and showing result improvements [49], [50]. In our approach, we follow this line of research because it allows to consider MIL into an end-to-end trained model.

MIL paradigm assumes neither ordering nor dependency of instances within a bag. However that does not apply in our problem since the selection of an action to be part of a summary is highly dependent on the sequence of its events. For instance, a penalty or a free-kick are always preceded by a foul. More importantly, the permutation of events could completely change the meaning or the interest of an action.

For this reason, we argue that fully-connected layers as proposed by previous works are not completely suitable to capture this sequentiality. Recurrent neural networks are better suited to model such dependency. At the core of the LSTMs are memory cells which encode, at every time step, the knowledge of the inputs that have been observed up to that step. Therefore, we propose an LSTM network followed by a MIL Pooling to get the bag representation.

Each input sample is a sequence of event feature vectors representing a bag. The sequence of feature vectors is then fed to an LSTM, and the hidden state of the LSTM is given by:

$$h_t = LSTM(h_{t-1}, x_{e_n}) \quad (2)$$

where $LSTM(h_{t-1}, x_{e_n})$ represents an LSTM function of hidden state h_{t-1} and input vector x_{e_n}.

Let b_f be a bag (of events) of size K. Each event is defined by its feature vector x_{e_n} (the detail of the event metadata features used in this work are provided in the supplementary material). We then learn an embedding for each event feature vector using an LSTM to preserve the sequential dependency between events. Let $H^{b_f} = \{h_1, ..., h_k\}$ be the K embeddings of the K events from bag b_f. Each h_k embedding is of size L. We propose then a MIL pooling schema to learn the final bag representation z^{b_f}, as defined in Eq.(3):

$$\forall_{l=1,...,L} : z_l^{b_f} = \max_{k=1,...,K} h_{kl} \quad (3)$$

where z^{b_f} is a feature vector of size L, and is obtained from getting the maximum of each position l across all the K event embeddings of the bag.

Finally, this representation z^{b_f} is input into a single sigmoid neuron which provides the score O_{b_f}, a value between 0 and 1, for the bag b_f to be a proposal or not.

B. Proposal Definition

From the output of the MIL network we get a score O_b providing the likelihood per bag to be proposed for the final summary, but the overlap between consecutive positions of the sliding window lead some events to belong to more than one bag and thus to get related to different scores O_b. In order to rely on the importance of each event to select or not an action in the final summary, we need to define a score per event instead of per bag, merging all the possible scores associated to the given event. The score S_{e_n} for the event e_n is then given by the Log-Sum-Exp (LSE) used in [48], the function is defined in Eq. (4). The LSE is a smooth version and convex approximation of the max function. The hyperparameter r controls the smoothness of approximation [49].

$$S_{e_n} \geq r^{-1} \cdot log \left[\frac{1}{|\{b_f \mid e_n \in b_f\}|} \sum_{b_f|e_n \in b_f} r \cdot O_{b_f} \right] \tag{4}$$

After obtaining the score per event, we use a threshold to select the positive events, this threshold is defined using the validation set. Thus an action proposal a_p is a set of positive consecutive events. We denote A the set of all action proposals in a match.

IV. Summarization: Hierarchical Multimodal Attention

One of the biggest challenges for automatic soccer video summarization is to produce summaries provoking as much emotion as the ones made by human operators. To decide which actions are added to the summary, human editors use different sources of information. For this reason, we propose a multimodal approach that use event metadata and audio.

A. Multimodality

The event data have been manually collected by human observers sitting inside the stadium during the game. Each time an event occurs on the field, the human annotates the event with: the type (e.g., pass, foul, out or card), a timestamp, the team and players involved, outcome (i.e., if the action has been successful), the location (i.e., (x,y) position) on the field. Depending on the type of event, other information is available. For example, the final position on the field of the event, and descriptors of each type of event (e.g., yellow or red for the event type card), which are called qualifiers. The detailed description of the metadata extracted from the event data is presented in the supplementary material.

Audio plays a very important role in sports, where crowd cheering and excitement in the commentators' voice are usually indicators of an important action. For this reason, our summarization approach does not only exploit the event metadata features but also the audio signal extracted from the broadcasted match video. We use 9 different audio features also detailed in the supplementary material. These features extract the energy, the spectrum, the MFCC and the variation of the audio signal.

Since each event e_n has a timestamp corresponding to a time in the match, we extract its corresponding video time $time_{e_n}$ in seconds. Since the different sounds in a broadcasted video come from the reactions of humans (spectators or commentators) just after the event occurs, the audio features of the event e_n are extracted from the interval $[time_{e_n}, time_{e_n} + 2]$. Inside this interval, the audio signal is first divided into short-term windows (frames) of 100 ms with 50% overlap, then for each frame all features are calculated.

B. Multimodal Attention

In Section III we have described how we use Multiple Instance Learning to obtain a score per event and how we have defined that an action proposal a_p is a set of positive consecutive events. Now the goal of the second block of our approach is to define which of these proposals indeed belong to the summary.

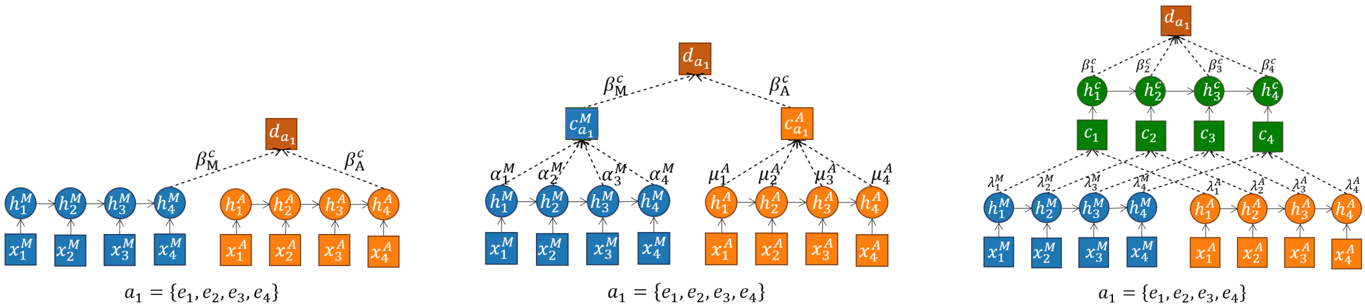
As we stated before, the use of multiple modalities is relevant to decide which actions are important in a sport match. The audio of an action can significantly vary not only from the type of the action but also from the events occurring inside the action. For instance, it is not the same kind of *goal* event, if the goal is preceded by several slow passes as if it is the result of an action starting by an interception or an error from the opponent team. For this reason, instead of learning the importance of each modality per action, we propose a hierarchical multimodal attention mechanism that in the first stage learns the importance of each modality at the event level and in the second stage learns the importance of each event inside the action (see Figure 2c).

Thus, in the first stage of the hierarchy, the multimodal representation vector per event is given by a weighted average:

$$c_i = \lambda_i^M h_i^M + \lambda_i^A h_i^A \tag{5}$$

where the weights of each modality $\{\lambda_i^M, \lambda_i^A\}$ are determined by an attention layer shared across time-steps, see Fig.2.

An action might contain several events that are not considered as important in a match but they are relevant to provide a context to the fans. For instance, there are many fouls during the match, but if a card action belongs to the summary it is important to show the foul and pass events that provoked this card. However, depending on the excitement of the crowd or the type of card, the importance of the events may vary. Therefore, after obtaining a multimodal representation per event c_i, we train an attention layer that learns the importance of each event inside the action, resulting in the weight β_i^c.


d_{a_1}
β^c_M
β^c_A
h^M_1 h^M_2 h^M_3 h^M_4
h^A_1 h^A_2 h^A_3 h^A_4
x^M_1 x^M_2 x^M_3 x^M_4
x^A_1 x^A_2 x^A_3 x^A_4
$a_1 = \{e_1, e_2, e_3, e_4\}$
$c^M_{a_1}$
$c^A_{a_1}$
α^M_1 α^M_2 α^M_3 α^M_4
μ^A_1 μ^A_2 μ^A_3 μ^A_4
β^c_1 β^c_2 β^c_3 β^c_4
h^c_1 h^c_2 h^c_3 h^c_4
c_1 c_2 c_3 c_4
λ^M_1 λ^M_2 λ^M_3 λ^M_4
λ^A_1 λ^A_2 λ^A_3 λ^A_4

(a) Naive baseline

(b) Hori et al. model

(c) Our model

Fig. 2: Definition of our hierarchical multimodal attention schema and comparison with state-of-the-art approaches. Blue indicates metadatada, orange indicates audio and green indicate the multimodal representation of the events. λ and β are attention weights.

Thus, the representation vector per action proposal is given by a weighted average:

$$d_{a_p} = \sum_{i=1}^{L_p} \beta^c_i h^c_i \tag{6}$$

where L_p is the length (number of events) of action a_p. Finally, each of this d_{a_p} action representation is given to a sigmoid neuron which outputs a value between 0 and 1, that indicates the likelihood of the action a_p to be included in the summary.

For the sake of space, we detail in the supplementary material the methodological comparison of our hierarchical multimodal attention model Fig.2c with a Naive baseline Fig.2a, and with Hori et al. model [50], Fig.2b.

V. Experiments

We first describe the dataset, then we split the evaluation in the two main tasks of our approach, the action proposals generation using LSTM MIL Pooling and the Summarization using the multimodal attention.

A. Dataset and metrics

Our dataset consists of event data for 20 matches from the 2017-2018 season of the French Ligue 1 and 50 matches from the 2019-2020 season of the English Premier League. There are 43 different types of events, related either to the flow of the match like a yellow card or to the action on the pitch like a shot. Each match corresponds to an average of 1700 events.

The only ground truth available are the 70 video summaries created by professional broadcasters. We created a set of intervals I^t, corresponding to the time location on the video match of all clips inside the summary. To obtain the ground truth in terms of events, we found all events that have a timestamp inside the intervals I^t and created a new set of intervals I. All the results reported in this section are evaluated with respect to the event ground truth I. The *Missing Intervals rate* and the F-score are computed from the fact that an interval refers to a summary interval of the video summaries created by professional broadcasters.

In order to obtain a fair comparison we use a 10-fold-cross-validation. Each fold has 80%, 10% and 10% of the matches for train, validation and test set respectively. The matches of the two leagues are equally distributed in each fold.

For all the experiments, we replicate the models from the literature using Keras library and choose the parameters that have shown high classification performance on our dataset.

B. Proposal Generation

We compare our LSTM MIL Pooling method with three state-of-the-art methods. For all the experiments reported on Proposal Generation, in test phase we consider that the method found an interval if at least 50% of the action is inside any interval of I.

As we mentioned previously, action proposal generation problem has been tackled by several approaches in video processing. SST [46] was created to generate temporal action proposals in untrimmed video sequences. It uses a recurrent neural network to produce confidence scores of multiple proposal sizes at each time step. We use the same approach but instead of video features, we use our metadata features as input. We use an LSTM network with 128 neurons, the output proposals sizes are $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ and in the training phase an output interval is considered as positive if at least 50% of it belongs to a candidate.

In terms of Multiple Instance Learning approaches, we compare with two Neural Networks based approaches. MI-Net [49] uses three fully connected layers to generate a representation per sample and then with a max-pooling layer over all the samples of the bag gets a score per bag. Ilse et al. [51] is a similar approach but instead of using a max-pooling layer, it adds an attention mechanism. For this reason in our experiments we call this method MI-Net Attention. The number of neurons for the three fully connected layers are 256, 128 and 64 respectively. For the attention we used 32 neurons.

For LSTM MIL Pooling we used a LSTM with 64 neurons, Adam optimizer, binary cross-entropy as loss function and a batch size of 32 bags.

For the MIL approaches, in the training phase a bag is considered positive if at least 50% of it belongs to a candidate.

As the goal of the first block of our method is to detect all the possible actions of the match, we will focus on a

TABLE I: Performance Comparison of Proposal Generation methods.

Method	Missing Intervals		Recall	
	Target Encoding	One-Hot Encoding	Target Encoding	One-Hot Encoding
SST [46]	39.79	45.53	60.11	54.35
MI-Net [49]	18.62	23.34	81.33	76.6
MI-Net Attention [51]	16.07	19.39	83.89	80.56
LSTM MIL Pooling	13.01	22.83	86.96	77.11

TABLE II: Performance comparison of Multimodal Attention methods.

Method	Missing Intervals	F-score
Sanabria et al. [48]	47.95	64.30
Naive Fusion	36.19	71.23
Hori et al. [50]	32.99	72.03
Ours	27.38	74.09

low Missing Intervals rate and a high Recall. This means we will pay more attention on not losing any potential summary interval of the match.

For Metadata we have both categorical and real-valued features. We consider also important to analyze the impact of mixing these types of features. Thus we compare the performance of the methods using two different representations for the categorical features, by converting the features either to One-Hot Encoding vectors or to real values using Target Encoding.

In One-Hot Encoding we created a vector of the total number of types/qualifiers with zero in all the positions and one in the position corresponding to the type/qualifier of the current event. On the other hand, Target Encoding gives only one real value that represents the fraction of times an event with this feature (type/qualifier) is labeled as 1 out of all the times an event with this feature is in the training set.

Table I shows that Target Encoding outperforms One-Hot Encoding in all the methods, which might be due to the sparsity of the features. There are 43 types of events and 47 qualifiers, an event can only have one event type and very rarely more than 3 qualifiers.

In Table I we can also see that MIL methods are clearly better on identifying the different proposals of the match, since SST performs at least 20% worse than the rest of the methods. LSTM MIL Pooling outperforms state-of-the art methods, it misses at least 3% less intervals and gets a Recall at least 3% higher compared with the second-best method (MI-Net Attention).

C. Summarization

The inputs of our Summarization block are the actions generated by the LSTM MIL Poolling method. We compare our Hierarchical Multimodal Attention model (cf. Fig.2c) with two state-of-the-art methods, Sanabria et al. [48] and Hori et al. [50] (see Fig.2b), and a baseline (see Fig.2a).

Although Sanabria et al. [48] method does not integrate an attention model, we consider it as a relevant approach to compare with, since their goal is to summarize soccer matches using event data. They propose to concatenate audio energy with metadata features to train a hierarchical LSTM to decide which actions belong to the summary. We used the same features and parameters as mentioned in their paper.

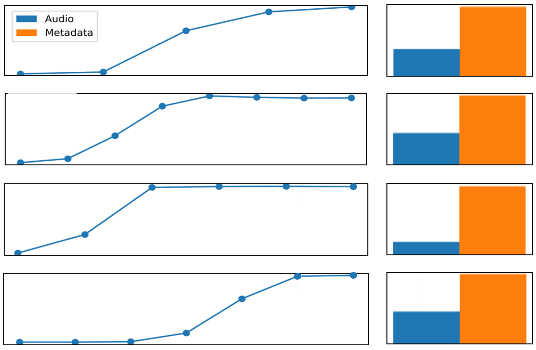


Fig. 3: Examples of attention in different actions of Hori et al [50]. On the left side, attention weights of audio part: The x axis is the sequence of events in the action and the y axis represents the weight values learned by the attention layer. On the right side, multimodal attention weights in the action level: the y axis is the weight values learned by the attention layer. Blue and orange represent the audio and the event metadata respectively.

As an attention baseline we have created a model called *Naive Fusion* (Fig.2a) that has a LSTM per modality, then the last state of each LSTM passes through an attention layer that learns the importance of each modality. Finally the weighted sum is passed to a sigmoid neuron to make the decision.

Hori et al. [50] proposed an attention-based multimodal fusion for video description (Fig.2b). We keep the same schema and replicate it for our video summarization problem.

For the Naive Fusion and Hori et al. approaches, we used 32 neurons for the LSTM of each modality and our audio and metadata features as input. Our model has 32 neurons in h^M and h^A, and 16 neurons in h^c

Table II shows the Missing Intervals rate and the F-score of the aforementioned methods. The big gap between the performance of Sanabria et al. method and the others shows that the concatenation of different modalities is not enough to learn a representation of the actions. In Table II we can also see that our method misses at least 5% less actions and gets an increase of 2% in F-score than the state-of-the-art method.

We believe that our method outperforms [50] because in this method the multimodal fusion is done at the action level. Indeed their method has an attention layer at event level but it is done separately per modality. Learning the importance of the event using only the audio features of a soccer match is a very difficult task (see sumpplementary material for a clearer explanation of the architectures). The left side of Figure 3 displays the attention learned in the audio part by [50] in four different actions. It seems that the attention is just learning that the last events (i.e. the end of the actions) are more important no matter the type of the events. The right side of this figure displays the importance learned by the attention for the audio

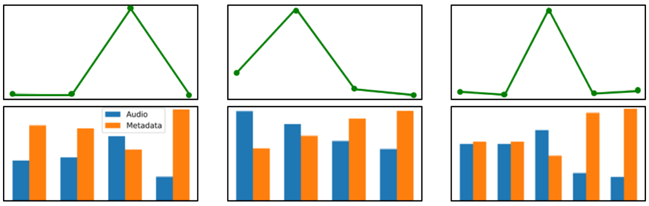


Fig. 4: Examples of attention in different actions learned by our model. On the bottom, multimodal attention weights at the event level: The x axis is the sequence of events in the action and the y axis represents the weight values learned by the attention layer. On the top, attention weights learned from the multimodal representation of each event. Blue and orange represent the audio and the metadata respectively.

TABLE III: Performance comparison of Soccer Baselines.

Method	Precision	Recall	F-score
Only Goals	99.55	28.29	44.18
All Shots-on-Target	40.77	75.71	52.99
Random	41.87	48.72	45.03
Ours	75.46	72.76	74.09

and metadata modalities. This not only shows that for this model the metadata is often more important but also that the audio modality is most of the times neglected.

On the other hand, Figure 4 shows some qualitative results of our model. We can see that the multimodal attention does not follow a particular pattern, audio and metadata importance can be very different from one action to another. And the attention learned by the second stage of our model considers many important events where the audio was considered as more relevant from the previous stage.

Soccer Baselines. As we previously mentioned, the evaluation of most of the methods on video sports summarization are based on the detection of most important actions, then to perform a fair comparison, we propose three baselines:

- *Only Goals*: Only the goals of the match are predicted as positive. Since the easiest way to create a summary from a soccer video is to extract the goals of the match.
- *All Shots-on-Target*: All Shots on Target actions (i.e. goals, goalkeeper saving a shot on goal, any shot on goal which goes wide or over the goal and whenever the ball hits the frame of the goal) are predicted as positive.
- *Random*: The prediction is a random value between 0 and 1, where the samples with values below 0.5 are negatives and the ones greater or equal than 0.5 are positives.

Table III depicts the performance of these baselines. Our F-score is clearly the highest. The Precision of our approach is only outperformed by *Only Goals*, considering it is very likely that all the goals of the match belong to the summary, however the Recall of this baseline is the lowest since it misses many other type of actions. The Recall of our approach is only outperformed by *All Shots-on-Target*, since the Shots on Target actions represent a big percentage of the actions included in summaries, yet the Precision of this baseline is at least 34% lower than ours.

TABLE IV: Performance comparison of Separete Modalities.

Method	Missing Intervals	F-score
Only Audio	27.75	64.17
Only Metadata	32.09	72.05
Ours	27.38	74.09

TABLE V: F1-score comparison of Audio Features.

Audio Features	No Attention	Attention
Energy Features	59.37	59.51
Our Features set	62.78	64.17

Multimodality. In order to show the importance of merging multiple modalities, we evaluate the performance of each modality separately using a LSTM with an attention layer.

Table IV shows that our method obtains the highest F-score compared with the models using only audio and only metadata features. Although using only audio features less intervals are missing, the low F-score reveals the low precision of this method since it predicts a lot of false positives. Comparing our results with the method using only the metadata we can see that adding the audio features helps to reduce almost 5% of missing intervals.

Audio Features. We consider that it is also important to show if the use of additional audio features improves the results. Sanabria et al. [48] only used the energy of the audio signal, however as it was mentioned in section IV-A, there are many other audio features that have helped to improve classifications in other contexts. We created two different models that take as input either the energy features proposed by Sanabria et al. or the audio features proposed in this paper (details are in the supplementary material), in order to predict which action belong to the summary. One model (*Attention* in Table V) has an attention layer that learns the importance of each event and the second model is a regular LSTM (*No-Attention* in Table V).

Table V shows that the models using our feature set outperforms at least by 2% the models using only the Energy features. And this behavior holds for models with and without attention mechanisms.

VI. Conclusion

In this paper, we proposed a method to summarize soccer matches using multimodal event data. We introduce a Multiple Instance Learning approach where the instances are sequentially dependent and we have empirically shown that it is a good method to generate proposals in a context where the instance labels are not available. Different from existing multimodal attention approaches, our method focuses on the importance of each modality at event level to then, in a second stage, learn the relevance of each event at action level. Experiments in a dataset composed of two different soccer leagues show the capability of our approach to identify the best actions for automatic summarization by comparing with real summaries made by human operators and outperforming state-of-the-art methods.

References

[1] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Identifying team style in soccer using formations learned from spatiotemporal tracking data," in *IEEE ICDM Workshops*, 2014, pp. 9–14.

[2] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews, "Assessing team strategy using spatiotemporal data," in *ACM SIGKDD*, 2013, pp. 1366–1374.

[3] J. Van Haaren, V. Dzyuba, S. Hannosset, and J. Davis, "Automatically discovering offensive patterns in soccer match data," in *International Symposium on Intelligent Data Analysis*. Springer, 2015, pp. 286–297.

[4] L. Gyarmati and X. Anguera, "Automatic extraction of the passing strategies of soccer teams," *arXiv preprint arXiv:1508.02171*, 2015.

[5] V. Vercruyssen, L. De Raedt, and J. Davis, "Qualitative spatial reasoning for soccer pass prediction," in *CEUR Workshop*, vol. 1842, 2016.

[6] T. Decroos, J. Van Haaren, and J. Davis, "Automatic discovery of tactics in spatio-temporal soccer match data," in *ACM SIGKDD*. ACM, 2018, pp. 223–232.

[7] G. Liu and O. Schulte, "Deep reinforcement learning in ice hockey for context-aware player evaluation," *arXiv preprint arXiv:1805.11088*, 2018.

[8] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, "Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach," *ACM TIST*, vol. 10, no. 5, pp. 1–27, 2019.

[9] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," in *ACM SIGKDD*, 2019, pp. 1851–1861.

[10] L. Bransen and J. Van Haaren, "Measuring football players' on-the-ball contributions from passes during games," in *International Workshop on Machine Learning and Data Mining for Sports Analytics*. Springer, 2018, pp. 3–15.

[11] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," *Sci Data*, vol. 6, no. 1, pp. 1–15, 2019.

[12] T. Bergmann, S. Bunk, J. Eschrig, C. Hentschel, M. Knuth, H. Sack, and R. Schüler, "Linked soccer data." in *I-SEMANTICS (Posters & Demos)*. Citeseer, 2013, pp. 25–29.

[13] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. on IP*, vol. 12, no. 7, pp. 796–807, 2003.

[14] M. Y. Eldib, B. S. A. Zaid, H. M. Zawbaa, M. El-Zahar, and M. El-Saban, "Soccer video summarization using enhanced logo detection," in *IEEE ICIP*, 2009, pp. 4345–4348.

[15] N. Nguyen and A. Yoshitaka, "Soccer video summarization based on cinematography and motion analysis," in *IEEE MMSP*, 2014, pp. 1–6.

[16] M. Tavassolipour, M. Karimian, and S. Kasaei, "Event detection and summarization in soccer videos using bayesian network and copula," *IEEE Trans. on CSVT*, vol. 24, no. 2, pp. 291–304, 2014.

[17] T. Liu, Y. Lu, X. Lei, L. Zhang, H. Wang, W. Huang, and Z. Wang, "Soccer video event detection using 3d convolutional networks and shot boundary detection via deep feature distance," in *ICONIP*. Springer, 2017, pp. 440–449.

[18] R. Agyeman, R. Muhammad, and G. S. Choi, "Soccer video summarization using deep learning," in *IEEE MIPR*, 2019, pp. 270–273.

[19] A. Javed, A. Irtaza, Y. Khaliq, H. Malik, and M. T. Mahmood, "Replay and key-events detection for sports video summarization using confined elliptical local ternary patterns and extreme learning machine," *Applied Intelligence*, pp. 1–19, 2019.

[20] D. Corney, C. Martin, and A. Göker, "Two sides to every story: Subjective event summarization of sports events using twitter." in *SoMuS@ ICMR*. Citeseer, 2014.

[21] A. Tang and S. Boring, "# epicplay: Crowd-sourcing sports video highlights," in *SIGCHI*. ACM, 2012, pp. 1569–1572.

[22] Y. Huang, C. Shen, and T. Li, "Event summarization for sports games using twitter streams," *WWW*, vol. 21, no. 3, pp. 609–627, 2018.

[23] E. Mendi, H. B. Clemente, and C. Bayrak, "Sports video summarization based on motion analysis," *Computers & Electrical Engineering*, vol. 39, no. 3, pp. 790–796, 2013.

[24] K. Tang, Y. Bao, Z. Zhao, L. Zhu, Y. Lin, and Y. Peng, "Autohighlight: Automatic highlights detection and segmentation in soccer matches," in *IEEE Big Data*, 2018, pp. 4619–4624.

[25] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for tv baseball programs," in *ACM Multimedia*, 2000, pp. 105–115.

[26] A. Baijal, J. Cho, W. Lee, and B.-S. Ko, "Sports highlights generation bas ed on acoustic events detection: A rugby case study," in *IEEE ICCE*, 2015, pp. 20–23.

[27] V. Bettadapura, C. Pantofaru, and I. Essa, "Leveraging contextual cues for generating basketball highlights," in *ACM Multimedia*, 2016, pp. 908–917.

[28] M. Merler, D. Joshi, Q.-B. Nguyen, S. Hammer, J. Kent, J. R. Smith, and R. S. Feris, "Automatic curation of golf highlights using multimodal excitement features," in *IEEE CVPRW*. IEEE, 2017, pp. 57–65.

[29] A. Raventos, R. Quijada, L. Torres, and F. Tarrés, "Automatic summarization of soccer highlights using audio-visual descriptors," *SpringerPlus*, vol. 4, no. 1, p. 301, 2015.

[30] P. Shukla, H. Sadana, A. Bansal, D. Verma, C. Elmadjian, B. Raman, and M. Turk, "Automatic cricket highlight generation using event-driven and excitement-based features," in *IEEE CVPRW*, 2018, pp. 1800–1808.

[31] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *IEEE CVPR*, 2016, pp. 1059–1067.

[32] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *IEEE CVPR*, 2015, pp. 3584–3592.

[33] R. Panda and A. K. Roy-Chowdhury, "Collaborative summarization of topic-related videos," in *IEEE CVPR*, 2017.

[34] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proceedings of ECCV*, 2018, pp. 347–363.

[35] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *IEEE CVPR*, 2015, pp. 3090–3098.

[36] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. on IP*, vol. 26, no. 8, pp. 3652–3664, 2017.

[37] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *IEEE CVPR*, 2019, pp. 7902–7911.

[38] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, "Unsupervised video summarization with attentive conditional generative adversarial networks," in *ACM Multimedia*, 2019, pp. 2296–2304.

[39] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *ACM Multimedia*, 2017, pp. 863–871.

[40] J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan, "Stacked memory network for video summarization," in *ACM Multimedia*, 2019, pp. 836–844.

[41] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *ECCV*. Springer, 2016, pp. 766–782.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.

[43] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in *IEEE CVPR*, 2016, pp. 845–853.

[44] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *IEEE CVPR*, 2018, pp. 8971–8980.

[45] H. Xu, A. Das, and K. Saenko, "R-c3d: region convolutional 3d network for temporal activity detection," in *IEEE ICCV*, 2017, pp. 5794–5803.

[46] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "Sst: Single-stream temporal action proposals," in *IEEE CVPR*, 2017, pp. 2911–2920.

[47] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *BMVC*, 2017.

[48] M. Sanabria, F. Precioso, T. Menguy *et al.*, "A deep architecture for multimodal summarization of soccer games," in *ACMM MMSports'19*. ACM, 2019, pp. 16–24.

[49] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[50] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *IEEE ICCV*, 2017, pp. 4193–4202.

[51] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.