Profiling Actions for Sport Video Summarization: An attention signal analysis

Melissa Sanabria, Frederic Precioso, Thomas Menguy

Profiling Actions for Sport Video Summarization: An attention signal analysis

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

***Abstract*—Currently, in broadcast companies many human operators select which actions should belong to the summary based on multiple rules they have built upon their own experience using different sources of information. These rules define the different profiles of actions of interest that help the operator to generate better customized summaries. Most of these profiles do not directly rely on broadcast video content but rather exploit metadata describing the course of the match. In this paper, we show how the signals produced by the attention layer of a recurrent neural network can be seen as a learned representation of these action profiles and provide a new tool to support operators' work. The results in soccer matches show the capacity of our approach to transfer knowledge between datasets from different broadcasting companies, from different leagues, and the ability of the attention layer to learn meaningful action profiles.**

***Index Terms*—Video Summarization, Sports Video, Event stream data, Neural Network, User-Centric**

I. Introduction

Multimedia content summarization is present in an ever-increasing number of areas. Sports, and particularly soccer, is one of the sectors that has invested the most in video analysis field, due to the popularity of the sport but also to the recent increase of sport betting platforms. To provide soccer fans, and gamblers with as much information as possible, the need of techniques to extract key information and summarize the increasing number of games has intensified.

Broadcasting companies usually do not rely on automatic algorithms to provide their audience with the summary of a soccer game almost right after the end of the game, instead they mainly rely on human operators aided by algorithms. These operators use the video content to build up a summary but most of their work is based on event-metadata. Indeed, processing the broadcast video would not be enough since some content is not shown on tv or possibly not under an optimal view angle, and watching while processing at the same time the content from all the cameras would not be tractable. Therefore, human operators exploit event-metadata provided by companies like Prozone, GeniusSports, Opta, WyScout, and many others. These companies dispatch human observers to all the stadiums so that they collect on live the events happening on the fields. The metadata relate to all the events of the matches like pass, basket, shot, hit, head, free-kick, corner, foul, cards, etc.

The number of events in a match is significantly smaller than the number of frames. In the case of soccer, the match duration is around 90 minutes, at a rate of 25 frames per second (without sub-sampling) it corresponds to 135000 frames of at least 112x112 pixels. On the other hand, the same match corresponds to about 1500 events overall (without sub-sampling), plus each event is represented by few values like the type, location and the distance to the goal. Therefore, to build summaries, operators have designed handmade decision rules exploiting all the information hold by these event metadata, to produce the result as quickly as possible. These handmade rules are based on the type, the speed or the order of the events to determine different profiles of actions.

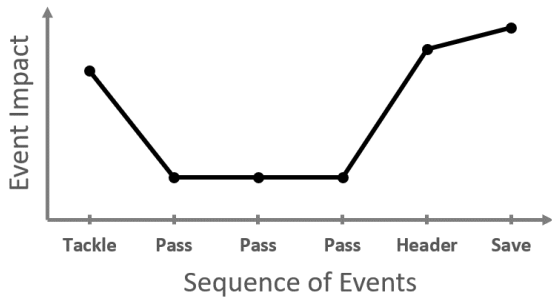

Fig. 1: Example of an action profile that an operator has in mind when selecting this action for the summary: the first tackle to get the ball was amazing corresponding to a high impact of the event in the overall interest of the action. Then two unexciting passes led to a final long assist that reached the striker. Despite being in the middle of the defense, he successfully headed the ball but it was unfortunately blocked by the goalkeeper.

How to make the choice between two *"shots on target"* (i.e. a shot that is very close to the goal frame but does not end in scoring a goal)? The operator looks at: How far from

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

the goalmouth the player shooting the ball was? From which angle? Who is the player shooting? At what time in the game the shoot happens? After which nice movement of the team? etc.

For instance in the above example (see Fig.1), the operator decides to choose this specific *"shot on target"* because before the goalkeeper caught the ball, the adversarial striker has made a very difficult header surrounded by adversaries after receiving the ball through a long pass from his winger, rewinding the action before this final assist there were two passes (without anything noticeable or emotional), but everything started with an amazing tackle to steal the ball. In this sequence of events, each event has a different weight representing the impact of the event's characteristics in the mind of the operator and so in the final choice of that action.

The question we intend to answer in this article is: Could we capture or learn in any way these mental representations leading an action to be relevant for the operator then added to the final summary?

Previous works have shown that internal signals produced by neural networks during training or inference provide meaningful information to understand their decision making process and to interpret which input parts are involved in the final decision [1], [2].

In this article, we are going to build an attentional recurrent neural network and show that its internal signals produce automatically from event-metadata the profile of each action, providing the operator with a relevant action representation to support the final decision. We show also that our method:

- generalizes from English Premier League to French Ligue 1 summaries using event-metadata
- generates meaningful profiles of the actions

II. Related Work

Early work in video summarization for sports mainly relies on hand-crafted heuristics [3]–[5], exploiting characteristics of the field or edition patterns like the replays. However, these methods show a clear lack of generalization.

Although a lot of methods try to solve the problem of video summarization for different sports, major obstacles remain: owing to copyright regulations, no benchmark datasets can be produced which makes more difficult the comparison among techniques; the objectives of video summarization themselves are often not well defined since many of the methods on video sport summarization do not evaluate using summarization metrics, they usually rather focus on the detection of most important actions like goals.

The field of video summarization of general-purpose content has progressed rapidly providing numerous approaches with deep learning techniques [6]–[8]. However, the objectives for general-purpose video summarization differ drastically from sport summarization. For instance, most of the methods target to maximize the diversity of the resulting summary while minimizing the number of similar shots [9]–[12]. Such an objective cannot apply for sports summarization. Indeed, for soccer the simplest summarization algorithm is to keep only goal clips which are of course all very similar visually, and this situation holds for many other sports and many other kinds of actions.

In terms of benchmark datasets for video summarization, differences are also striking between general-purpose videos and sport ones [13]–[17]. The length of general-purpose videos varies from 6 to 10 minutes and the summary length is about 15% of the original video while, for sport videos, a match can vary from one to several hours depending on the sport. This difference is as striking when it comes to the summaries since they are about few minutes, for instance 5 minutes (which is already a long summary for a soccer game) would represent less than 6% of the original video of a soccer game.

The question of video duration is not only a limitation to evaluate against existing benchmarks for general-purpose video summarization, but also in terms of computation. For instance, with input samples of at least one hour long, keeping only 3 frames per second as it is done in standard state-of-the-art works [18], would lead to 10800 frames per match and per camera. Indeed, the video frames of a broadcast match contain replays or advertisements which makes many of these frames useless for summarization. To solve this computational problem, several approaches have recently used event-metadata to extract important features from sport matches [19]–[22] and rely only on the information coming directly from the field itself without considering the heavily edited broadcast video.

These event-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) [23]–[25]. Furthermore, these event-metadata correspond to the event features that operators use to build their mental representation of events and actions that we have described in the introduction. Our approach will thus rely on these same event metadata.

III. Approach

We describe a match as a sequence of events $e_1, e_2, ..., e_n$, which is a record of all the actual events happening on the field including passes, shots, cards, fouls, etc. We denote actions in the match by a_k, referring to a consecutive subset of events that does not overlap with others. For instance a goal action a_3 might be composed of five consecutive events $a_3 = \{e_9, e_{10}, e_{11}, e_{12}, e_{13}\} = \{pass, pass, tackle, pass, goal\}$. In this formulation x_{e_j} represents the feature vector of the j^{th} event e_j (see Figure 2 left). This feature vector is very heterogeneous since it is made of real values for x and y positions on the field, binary value for the outcome of the event, an integer to identify the event type, etc. The precise description of the feature vectors will be provided in the section IV-A.

A. Model

Our model consists of an LSTM network with an attention layer that captures the importance of each time step.

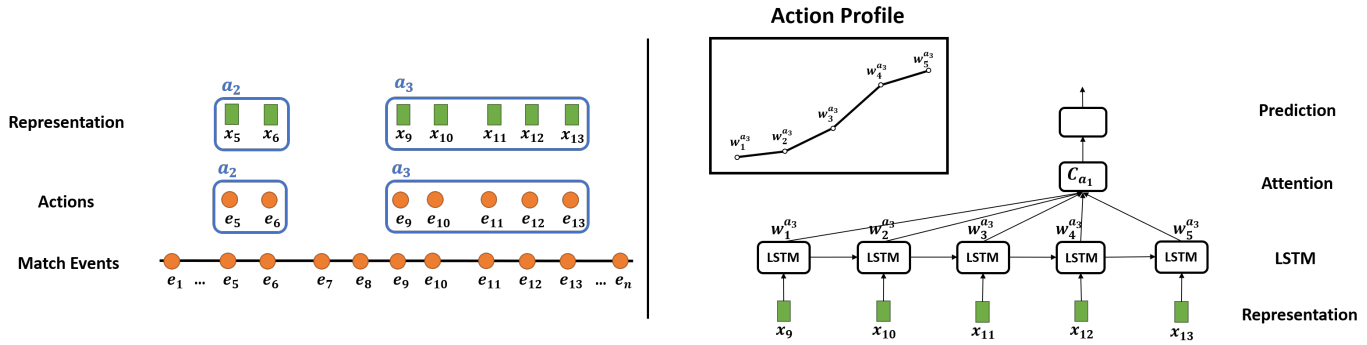

Fig. 2: How the action profiles are generated. On the left, the description of events and actions used in this paper. On the right, the LSTM model with an attention layer, showing on the top an example of our proposed graphical representation of an action profile.

Attentional models have been used before for translation, speech recognition and image caption generation [26]–[28]. Our attentional approach is like the image captioning methods [29] where the recurrent model learns to focus on the relevant parts of the image to better describe it. Our intuition is that this focus of attention is similar to the focus of attention of the operator.

In our approach, the LSTM takes as input the events of an action and predicts the likelihood of this action to be selected in the summary.

To be more precise let us take the action a_3 as an example (see Figure 2 right). This action has $L = 5$ events and the input of the LSTM is $\{x_9, x_{10}, x_{11}, x_{12}, x_{13}\}$. We denote $h_l^{a_3}$ as the hidden state of the LSTM unit at each time step l of action a_3. Then the attentional weights are defined as:

$$w^{a_3} = softmax(tanh(f(h_l^{a_3}))) \tag{1}$$

where f are the parameters of a single neuron.

Finally, the output of the final state of the LSTM is a weighted sum of all the hidden states of the action:

$$h_L^{a_3} = \sum_{l=1}^{L} w_l^{a_3} h_l^{a_3} \tag{2}$$

This final state is the input for a last sigmoid neuron that outputs a value between 0 and 1 that represents the likelihood of the action to be selected in the summary.

B. Graphical Action Profiles

Automatic summarization is important but sometimes the decision of whether an action is in the summary depends on different aspects like the style of editing, the enthusiasm of the fans or the target length of the resulting summary. Therefore, it is important to give additional information and different options to the operator.

Operators have usually designed hand-crafted rules to determine different profiles of the actions, these rules are based on the type, the speed or the order of the events. These profiles represent different options of the same type of action. For instance, two possible profiles for a Goal action might be: the first one including several events before the actual goal because it was a very quick action, and a second one including only two events before the goal because it was preceded by a free-kick.

We have the intuition that the attention layer of our approach can implicitly learn a representation of the action profile. This new representation provides a new tool for the operator that might help her/him taking decisions.

We propose to extract the weights learned by the attention layer and plot them in a graph, where the x-axis represents the sequence of events and the y-axis is the weight value (see Figure 3). Hence, we create a graphical representation of the action profile.

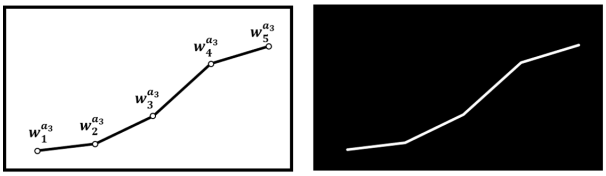

Fig. 3: Graphical Action Profiles. This is an example for an action composed by five events. On the left, it is the curve generated from the weights learned by the attention layer, the x-axis represents the event order and the y-axis is the weight value. And on the right, it is the image representation used for the classification task.

IV. Experiments

We first introduce the experimental setting, describing the datasets and features. We then present the capacity of our approach to transfer knowledge to other soccer leagues and the ability of the attention layer to generate meaningful profiles.

A. Setup

Dataset. We consider 70 matches from two soccer leagues: 50 matches of the English Premier League for the 2019-2020

season, and 20 matches of the French Ligue 1 for the 2017-2018 season. We have used a nested cross validation technique with 10 folds where each fold has 80%, 10% and 10% of matches for train, test and validation respectively.

The only ground truth available are the 70 video summaries provided by professional broadcasters. In order to obtain the actions of the matches which could be candidates to be in the final summary, we first list all the sequences of events found in the ground truth summaries made by the operators for all the videos in the dataset. Then we search all the instances of the same sequences of events in all the videos. For example, in one of the summaries created by an operator, there is a goal clip which corresponds to the sequence of events: long ball, aerial, pass, goal. Thus, we look for this exact same sequence in the rest of the match and in all the remaining matches in the training set to annotate them as ground-truth candidate actions. Any action that overlaps with any ground-truth sequence is labeled as positive and negative otherwise. It is important to notice that this is just a way to obtain the candidate actions of a match, we could use any other method to extract them.

LSTM Features. Our data consist of event-metadata that were manually collected by human observers located in the stadium, which in a real life context is provided by several companies such as Prozone, GeniusSports, Opta, WyScout, and many others. Each time an event happens on the field, the human observer annotates the event with, the location, timestamp, type (e.g., pass, card, out or miss) and the players who are involved. Depending on the type of the event, additional information is available like, the outcome of the action (e.g., if the pass was received by a player of the same team the outcome is 1 and if it is intercepted by an opponent the outcome is 0) and the qualifiers that describe the events (e.g., long ball, red card or head pass). For our experiments, we use the x and y positions on the field where the event started and ended, 0 and 1 values for the outcome of the action, a target encoding representation for the event type, another target encoding representation for the qualifiers, the elapsed time between two consecutive events, the distances to the goal for the event's start and end locations, and the angles to the goal for the event's start and end locations.

Profile Features. The images of the profiles have black background and the curve representing the attention weights is white, as shown in the right side of Figure 3. We further extract GoogleNet features from these images in order to take the final decision: "should be in the summary or not?".

B. Results

As aforementioned, our main goal in this work is to produce meaningful action profiles to help operators in selecting as quickly as possible the right sequences of events to put in the summary. In order to evaluate the quality of the profiles our attention LSTM has produced, without requesting the feedback from human operators, we train an SVM to determine whether the profile learned by the attentional layer indeed corresponds to an action that belongs to the summary. We generate an image profile for each of the actions of the dataset, then extract GoogleNet features from each of these images and train an SVM. The ground-truth is the same as the one used for the LSTM.

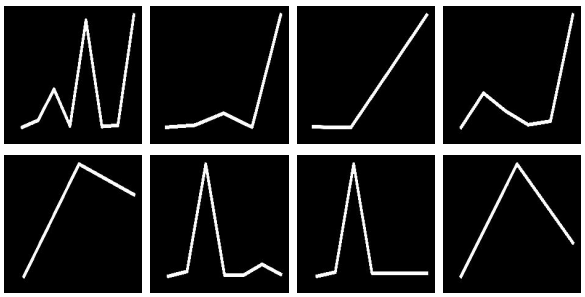

Fig. 4: Examples of profiles for *Goal* actions (top) and *Miss* actions (bottom).

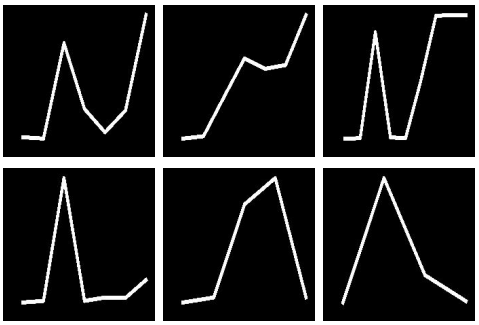

Fig. 5: Profiles of positive (top) and negative (bottom) actions.

Some examples of the image profiles are shown in Figure 4. Four different *Goal* action profiles are depicted on the top of the Figure, where we clearly can see that the attention layer learned that the last event was very important, this last event is the actual Goal event. On the bottom of this figure, there are four examples of *Miss* action profiles. In Figure 5 we also can differentiate the profiles from positive and negative, i.e. actions that belong to the summary (top of the Figure) and actions that do not belong to the summary (bottom of the Figure).

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 I depicts the performance on our dataset. Our F1-score is clearly the highest. The Precision of our approach is only outperformed by *Only Goals*, considering it is very

Method	Precision	Recall	F1
Ours	87.06	72.29	78.86
Only Goals	99.55	26.94	42.23
All Shots-on-Target	39.74	76.18	52.15
Random	20.16	47.58	28.21

TABLE I: Classification results using graphical action profiles. *Only Goals* predicts all the goals as part of the summary. *All Shots-on-Target* predicts all the Shots on Target as part of the summary.

Method	Precision	Recall	F1
Ours	79.88	68.01	72.73
Only Goals	100	31.93	47.14
All Shots-on-Target	37.72	73.84	49.44
Random	21.19	60.25	31.28

TABLE II: Generalization on classification results using graphical action profiles. All the results correspond to the classification scores for Ligue 1 matches. For *Ours*, the model was trained only with the Premier League matches. *Only Goals* predicts all the goals as part of the summary. *All Shots-on-Target* predicts all the Shots on Target as part of the summary.

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 47% lower than ours. With these results we show the relevance of the graphical representation of the action profiles learned by our attention layer to help determining if an action is a good candidate for the summary.

In order to prove the ability of our method to adapt properly to new unseen data, we also train our model using only the matches from the *Premier League* and analyze the results on matches from *Ligue 1*. The performance results on Table II shows the relevance and generalizability of the representation our model learns from the data.

V. CONCLUSION

In this article, we have explored how current learning models could support human operators who produce handmade summaries which are broadcast right after each (soccer) match. Based on the knowledge that these operators use decision rules built on event metadata to keep them or not in the summary, we have used an LSTM architecture model with an attention layer. The proposed approach generates images of action profiles from the weights learned by the attention layer. We prove the generalizability of our model which can learn from content provided by different broadcasters. We have also shown that the generated profiles contain meaningful information for the summarization task, since we can train an SVM to produce automatically a reasonably good summary from these profiles. Finally, neural networks produce internal signals (we could say intermediate signals), that are not only useful to understand better or to interpret the decision making process but these signals can also provide a new useful representation of the initial problem and lead to analyze it from a different perspective. In our future work, we will add more features on the event such as the identification of the player involved in doing the event because it seems very likely that an action involving one of the stars on the field would end in the summary. To evaluate this effect we need enough soccer matches with some of these stars playing, so we are currently focusing on collecting these games specifically. We would also evaluate how much our approach can be adjusted to one particular operator, since of course there are some differences between two operators summarizing the same game. To do so we will need enough matches summarized by the operator A, and enough matches summarized by the operator B, to see if we can learn the peculiarities and differences in their respective mental representations.

REFERENCES

[1] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

[2] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[3] Ahmet Ekin, A Murat Tekalp, and Rajiv Mehrotra,"Automatic soccer video analysis and summarization" IEEE Transactions on Image processing, vol. 12, no. 7,pp. 796–807, 2003.

[4] Mohamed Y Eldib, Bassam S Abou Zaid, Hossam MZawbaa, Mohamed El-Zahar, and Motaz El-Saban, "Soccer video summarization using enhanced logo de-tection," in 2009 16th ICIP. IEEE, 2009, pp. 4345–4348.

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

[6] Ke Zhang, Kristen Grauman, and Fei Sha, "Retrospective encoders for video summarization" in ECCV, 2018,pp. 383–399.

[7] Bin Zhao, Xuelong Li, and Xiaoqiang Lu, "Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization" in Proceedings of the IEEE CVPR, 2018, pp.7405–7414.

[8] Mrigank Rochan, Linwei Ye, and Yang Wang, "Video summarization using fully convolutional sequence networks," in Proceedings of ECCV, 2018, pp. 347–363.

[9] Junbo Wang, Wei Wang, Zhiyong Wang, Liang Wang,Dagan Feng, and Tieniu Tan, "Stacked memory network for video summarization," in Proceedings of the 27th ACM MM. ACM, 2019, pp. 836–844.

[10] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman, "Video summarization with long short-term memory," in ECCV. Springer, 2016, pp.

[11] Mrigank Rochan and Yang Wang, "Video summarization by learning from unpaired data" in Proceedings of the IEEE CVPR, 2019, pp. 7902–7911.

[12] Xuelong Li, Bin Zhao, and Xiaoqiang Lu, "A general framework for edited video and raw video summarization" IEEE Transactions on Image Processing, vol. 26,no. 8, pp. 3652–3664, 2017.

[13] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool, "Creating summaries from user videos" in ECCV. Springer, 2014, pp. 505–520.

[14] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes, "Tvsum: Summarizing web videos using titles" in Proceedings of the IEEE CVPR, 2015, pp.5179–5187.

[15] Sandra Eliza Fontes De Avila, Ana Paula Brand Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araujo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method" Pattern Recognition Letters, vol. 32, no. 1, pp.56–68, 2011.

[16] OVP 2011, "Open video project" 2011.

[17] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun, "Generation for user generated videos" in ECCV. Springer, 2016, pp. 609–625.
[18] Huijuan Xu, Abir Das, and Kate Saenko, "R-c3d: region convolutional 3d network for temporal activity detection" in ICCV,2017, pp. 5794–5803.
[19] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis, "Actions speak louder than goals: Valuing player actions in soccer" in Proceedings of the 25th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 1851–1861.
[20] Guiliang Liu and Oliver Schulte, "Deep reinforcement learning in ice hockey for context-aware player evaluation" arXiv preprint arXiv:1805.11088, 2018.
[21] Tom Decroos, Vladimir Dzyuba, Jan Van Haaren, andJesse Davis, "Predicting soccer highlights from spatio-temporal match event streams" in Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[22] Lotte Bransen and Jan 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.
[23] Pappalardo, Luca, et al. "A public data set of spatio-temporal match events in soccer competitions." Scientific data 6.1 (2019): 1-15.
[24] Mathien, H.: European Soccer Database. (2016) Data retrieved from http://www.kaggle.com/hugomathien/soccer
[25] Bergmann, Tanja, et al. "Linked Soccer Data." I-SEMANTICS (Posters and Demos). 2013.
[26] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).
[27] Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. "Automatic speech emotion recognition using recurrent neural networks with local attention." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.
[28] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.
[29] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho,Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in ICML, 2015, pp. 2048–2057