State of the Art - An Emotional Sports Highlight Generator*

Margaux Cavagna[1], Ahmed Abdel Rahman[1], Robert Cinciuc[1], and Bruno Sader[1]

INSA de Lyon, Computer Science Department, 69100 Villeurbanne, France

Abstract. Soccer highlight extractors are more and more frequent in this day and age. Most, if not all, focus on key moments, whether they be goals, bookings or substitute and use external sources to accurately curate them. In our paper, we study the usage of multi-modal models applied to TV streams in order to automatically generate highlights based on "emotional" events.
We will compare different approaches, for each mode (be it Audio or Video), study models used in different sports and determine which pair would work best for our use case. In order to score each model, we will evaluate them on :

- Ease of implementation
- Generalisability and transposability
- Accuracy

The study of the state of the art will be used in our next paper as a guideline.

Keywords: Machine Learning · Emotions Detector · Audio · Computer Vision.

1 Introduction

We are studying a multi-modal approach to generate soccer highlights. Most models focus on *key moment extractions* but we are looking into creating an *emotional moment extractor* in order to make people want to watch soccer. This paper is organized as follows. In Sect. 2, application of audio stream features is reviewed. Sect. 3 reviews different approaches to extracting information from video streams. In the last section we discuss how extracting available metadata could give extra insight on the segment.

2 Audio analysis

Audio is an excellent medium for capturing metadata about any activity presented, in our case it being soccer matches. Extraction of information from the

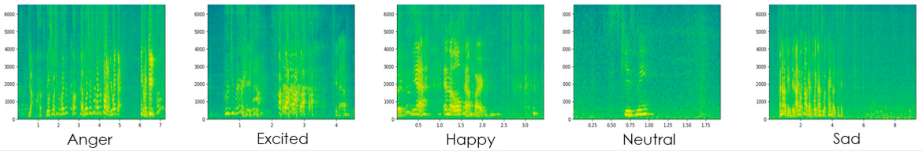
* Supported by Ignifai, Liris and Icar.

audio source could be done by 2 ways : analyzing the audio characteristics themselves or extracting the text from the speech of commentators and performing text mining. The first method provides features such as the pitch of sound, its volume, variation in intensity as well as the ambient noise. These characteristics inform us on the cheering of fans in the stands, the emotional state of the commentators, the game noises (whistle, players shouting) and unexpected events happening during the game (ads during the transmission, breaks from the actual game). Meanwhile, the second approach gives us insights on the action happening and allows us to classify sequences and search for important terms later on.

2.1 Audio features exploitation

As a rule of thumb, the cleaner the input into a system, the more chance there is for the results to hit the targeted values. Thus, the initial audio file needs to have as little noise as possible for the results to be of a high precision and recall. The more precise the searched class is, the bigger the noise set is in that case. For instance, if the algorithm tries to find only commentators' voices, all the other sounds may be treated as noise. Meanwhile, if we try to classify multiple types of sounds, our noise may be represented by the faults in the audio recording or external sounds such as birds singing. Noise removal is generally performed using a threshold to separate important audio sequences from the noise passages. One historical way of doing this seen in [1] is to divide the audio into small segments and randomly pick 10% of them, compute the average magnitude and take it as the threshold of noise. This method can, in some cases, give bad results due to the random nature of the algorithm. Another paper [2] presents a similar approach with a more robust technique. Using high sampling rates, read the audio file step by step and compute the maximum amplitude for each audio frame. The noise removal is done locally, based on a percentage of the maximum local amplitude, which has the advantage of adapting itself to each portion of audio. Knowing that sport matches have multiple adjacent segments with higher volume intensities (important points of a match), the threshold is updated to the corresponding volume.

The audio characteristics can be exploited by either analyzing directly the audio source or by creating a spectrogram of frequencies and training the algorithms on it. In order to create a spectrogram from a long audio sequence, one has to apply the short-term Fourier transformation (STFT) to obtain audio frames of equal length, then apply the fast Fourier transformation (FFT) to those which gives the Fourier spectrum of each frame [2]. The resulting 2D spectrogram has the time on the X axis, the range of frequencies on the Y axis and the magnitude of the amplitude's variation is depicted through the colors (dimmer colors represent smaller variation and brighter colors show high variation). An example of a spectrogram used in Speech Emotion Recognition (SER) can be seen in the figure 1. The spectrograms can be later used as input for training a neural network. An alternative to creating the spectrogram could be employing Mel Frequency Cepstral Coefficients (MFCC) and their first order differential coefficients as seen in [3], commonly known as delta-MFCC, to extract requisite features. Delta-MFCC captures the dynamics of the audio, such as the variation of MFCC coefficients with time.



[redacted]

Fig. 1: Visual representations of speech signal in 2D spectrograms of various emotions.

The next step in audio processing is labeling the segments by their corresponding semantic content. If the audio is used as the direct source of learning, it is divided into segments of similar characteristics based on the features learned from MFCC. The classes of sound are then modeled using either Gaussian Mixture Models (GMM) (defined in [3]) or Hidden Markov Model (HMM) (defined in [4]). The classification labels vary from basic ones such as : silence, applause and speech in [5], to more complex models: whistle, excited commentators' speech, unexcited speech, and music in [3]. More complex labeling systems can provide finer classifications, thus more variety in highlights generation, but the precision with such sets can be significantly worse compared to simple ones. Having the segments of fixed length and labels associated to them, we can proceed to the highlight generator part.

If the system uses spectrograms as the input for the algorithm, an adapted approach is the implementation of a neural network, such as Deep Stride CNN Architecture (DSCNN) [2]. Deep neural networks are excellent at identifying patterns in images and classifying them. This latter paper presents a neural network with only down sampling feature maps, instead of the usual pooling layers, which results in efficient computation. Because each emotional state affects the speech of a person, we can identify the state of the speaker by analyzing its spectrogram. Such a technique allows us to understand when the commentators are happy, sad, excited, neutral or angry. This provides further indicators to what segments of the match should be included as highlights of a game, because they wake a strong emotional response from the commentators.

If, on the other hand, direct audio is the input of the algorithm, multiple variants of the same principle exist for selecting important segments from those that present no subjective interest. The paper [4] presents an innovative approach to finding unusual key moments which replicates to some extent the way humans classify remarkable moments. The authors proposed a method using a sliding windows WL, composed of multiple WS windows. We compute the distribution of labels for each WS then separately for each position of WL, the latter shifting one WS at a time. A large histogram distance metric would indicate a rare or unusual event in the match. The maximum distance is associated with a WS window. Next, when shifting the WL again and detecting a certain threshold percentage distance of the maximum, a new highlight is said to start at the beginning of WL. The highlight continues until the histogram distance metric falls below the threshold. Then, the end of the highlight is the end of the WL window at the previous position. This way, if the bigger window indicates the predominant label is "cheering" while the smaller window says "excited speech", then an important moment should be in that time frame. The 2006 paper [5] by Zhang et al. introduces a completely new architecture, while having the same idea of 2 windows of different size the larger one being shifted along the audio. Researchers presented two architectures, the better performing being a hierarchical classifier. The cleaned audio passes through an MFCC Extractor and feeds the features to a first classifier (no explicitly mentioned type) focusing on short-term features and in parallel to a texture extractor, which is the equivalent of computing the long-term features of a WL window. Then, both outputs pass through a second classifier, thanks to which, both short-term features and long-term ones are considered during the classification. The last variation of the methodology we would like to present is the one the 2015 paper [3] by Baijal et al. Just as with other methods, Gaussian Mixture Model (GMM) is used to classify the segments. This time, the classification is multistage, which allows the algorithm to prioritize certain classes in favor of others. The authors of the paper first divide the segments into "speech" and "non-speech" categories. The first one is drilled further into "excited speech" and "unexcited speech". This way, we can make sure to capture the most important class, here being commentators' voices, while the other features remain less important for final highlight selection.

Finally, the selected segments must be polished before the users see them, to ensure a higher customer satisfaction. One technique found in [1] involves grouping continuous segments of important classes, such as cheering or applause and finding the maximum length of such a group. Then, a true highlight is considered a group with a longer length than a certain percentage threshold of the maximum length. To ensure all moments are included in the segment, the authors suggest including inside a group a number of frames before and after its time frame, the number of frames being subjectively chosen. Another paper [4] proposes the use of a Hidden Markov Model (HMM) to discriminate between actual interesting moments of a match and ads that may including high volume amplitude variations, which would result in a false positive in the first classification.

When it comes to results, we see a big discrepancy between older papers and newer ones. The works from 2003 [1] and 2004 [4] done in the same laboratory have quite low precision and recall compared to newer works. The first one has a precision varying between 40 and 60% with a very fluctuating recall. These results are caused by the goal of creating a unified classification model for multiple sports in parallel. The latter work improves on these results, but still doesn't perform on part with current techniques. The 2006 work in [5] has surprisingly high results for its time, achieving 98% precision and 96% recall. It's important to mention that this work focused on a simpler set of classes. The 2015 paper [3] achieved similar results as the previous work with an added plus of high user rating (4.23 on a 5 star scale) on rugby matches. Finally, the 2019 paper [2] focused on detecting human emotions achieved 81.75% prediction accuracy Interactive emotional dyadic motion capture (IEMOCAP) dataset and 79.5% for the Ryerson audio visual database of emotional speech and song (RAVDESS) dataset.

3 Video analysis

3.1 Scene detection

In this subsection, we will analyse scene changes. The goal is to recognize whenever a specific image shows up on screen. This could be for example, a quick video sequence containing the logo of the competition the match takes place in. Usually, sports channels will use such a logo whenever they are replaying an interesting event from the match. We would thus be interested in recognizing when this logo is shown, so we can use either the moment as a timestamp for after the action has happened, or we can use the replay that usually plays out after the logo sequence.

The other need for scene change analysis is analysing changes in the score bar. Matches will usually have a score bar in the upper left corner that gives us interesting information. First of all, the score bar gives us the time of the match, which is interesting data to get timestamps, for example. Obviously, the score bar gives us the score of the match. An easy way to see if a goal has been scored is to track a change on the scorebar. We can then rewind the video in order to get the goal footage. Finally, when the sports channel replays an action (such as a foul, an interesting move by one player, a goal...), the score bar disappears from the screen during the time of the replay. We can then combine this with the fact that a logo sequence will usually be played before and after the replay to either isolate the replay or once again, rewind the video to get the full on action from the beginning.

Therefore, scene change analysis is very important in our goal of extracting key moments from a football match automatically.

Recognising logo sequences. As stated above, it's very important to be able to recognize moments from the game in which the channel has shown a replay. The replay is composed of a logo sequence, as described above, and usually a slowed-down sequence of the interesting action. In order to recognize the replay section, Rehman and Saba in [8] use a technique called slow motion feature. The change of shots and its change frequency are detected and compared to the change frequencies of normal frames. This technique allows to analyse the video and extract all the replay moments easily.

Xu and al. in [9] also state that when using MPEG video formats, we can compute the numbers of both forward predicted and backward predicated macroblocks in B frames to determine whether or not a slow-motion replay is present.

Their paper also presents a variety of methods for specific video types such as using the transition effect of a specific DVE (Digital Video Effect), detecting gradual transitions and applying a Hidden Markov Model. Overall, we think that using a method that is applicable to a wide range of video formats is more efficient and easier to standardize, hence why we are focusing on logo transition, which today exists in all retransmitted professional matches.

The recognition of logo transitions is based on the fact that all those logo transitions have fixed duration, similar color histograms and similar images.

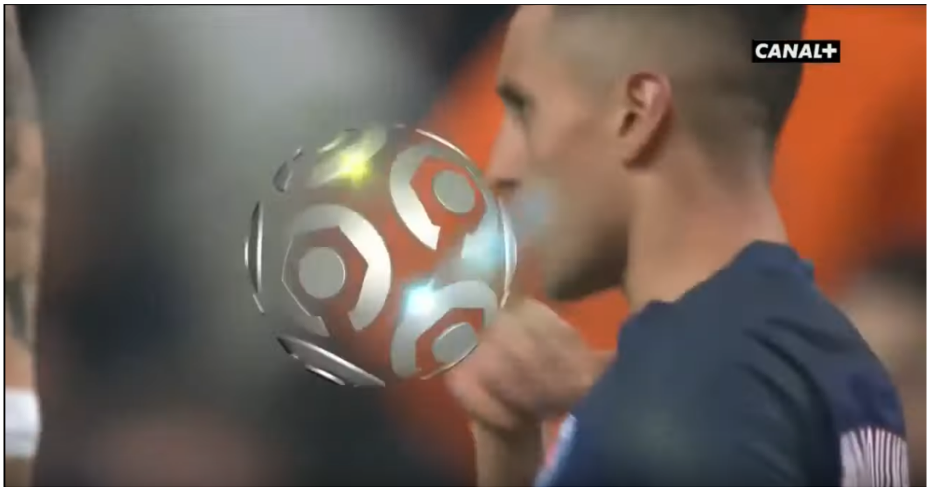


Fig. 2: Example of a logo transition : Canal +, Ligue 1 season 2016/2017. Whenever a sequence is replayed, a silver ball shows up on the screen in a sliding motion.

Xu and al. apply the following method to isolate these logo transitions, based on analysis of color histograms of the video : (a color histogram is a representation of the distribution of colors in an image)

1. Select the segments from the video stream that have a color histogram significantly different from the average color histogram of the stream. For example, we can imagine that during a football match, whenever we show the pitch there will be a large amount of green in the histogram, whereas logo transitions will have more “unusual” colors. They then recut all the segments into short clips based on the similarities of the color histograms.
2. Cluster said short clips in groups based on the average color histograms of the clips.
3. Define a threshold for image difference. Then, in each group, we calculate the image difference between the frames. If the number of small image differences is larger than the given threshold, the group is considered the logo group.

4. We can then detect logo transitions based on the logo color model of the logo group.

Recognising score bar. The task of recognizing a score bar has been done by Shukla and al. in their research [10]. Using an OCR (Optical Character Recognition) classifier, they locate the positions of the digits on the screen using bounding boxes. They train the classifier to recognize digits and dashes, which are present inside the scoreboard. They trained their classifier using an AlexNet model (one implementation of a convolutional neural network) that was trained on a multiclass linear SVM using a one-vs-all strategy (offset value = 0, C value = 1.0). They were then able to extract the digits from the scoreboard and to analyse them. We could apply this approach to our video analysis in order to extract interesting moments from the video.

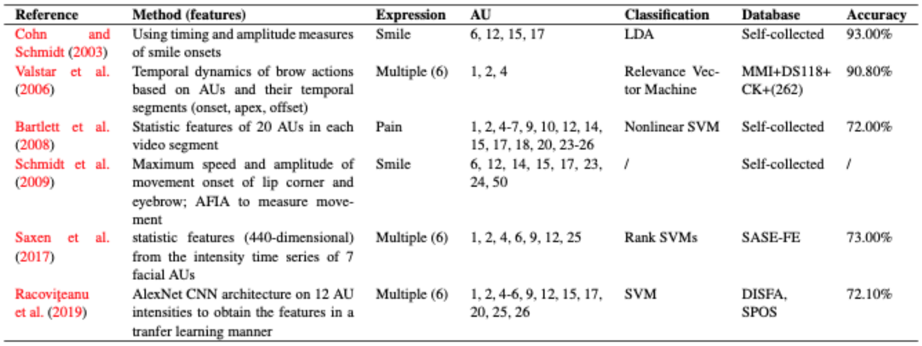
3.2 Extracting emotions from video streams

Each worthy moment in a match is followed by a change in scene and a close up of players, gathering, cheering, fighting etc... Many researches work on the analysis of emotions in videos. We will study how facial recognition and body language could help us evaluate in a close up shot how a player is reacting to an action and determine if the action was interesting. For instance, we could use a goal celebration as way to determine the emotion on the field (surprise, happiness, sadness ...) and combine this information with those extracted by the audio. This would allow us to more thoroughly extract goals as well as extract, if needed, only the goals we deem interesting (out of the ordinary).

Extracting emotions from facial expressions. Facial expressions have been studied for decades as a way to understand human emotions. Studies focus on 6 main emotions : 1. Anger 2. Happiness 3. Disgust 4. Sadness 5. Surprise 6. Contempt As explained in [11] facial expressions can be separated into two main classes, *Spontaneous* (which we are interested in) versus *Posed* (SVP). The study compared different models (see fig 3 and fig 4) and compared them on different categories. The ones we are interested in are *muscle movement (action units) based* and *hybrid methods*.

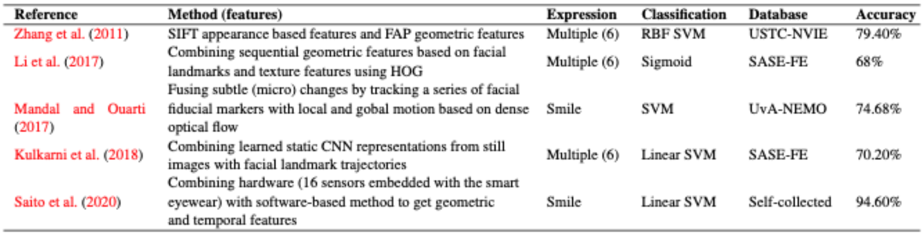
The study also goes in details on the influence of classifiers (which we'll come back to later) and determines that widely-used classifier SVM provides outstanding performances on several databases.

A recent study by Munasinghe [12] worked on extracting facial landmarks in order to recognise the facial emotion and compared different models on the CK+ database [17]. This paper goes into detail on how to implement facial detection using python. Landmark detection is used following the work done by Vahid et. al [13]. Considering 8 landmarks related to the eyebrows, 6 to the mouth and their normalised distance, we could create a feature vector. The author has compared different classification approaches, Random Forest, SVM, deep RNN



Reference	Method (features)	Expression	AU	Classification	Database	Accuracy
Cohn and Schmidt (2003)	Using timing and amplitude measures of smile onsets	Smile	6, 12, 15, 17	LDA	Self-collected	93.00%
Valstar et al. (2006)	Temporal dynamics of brow actions based on AUs and their temporal segments (onset, apex, offset)	Multiple (6)	1, 2, 4	Relevance Vector Machine	MMI+DS118+ CK+(262)	90.80%
Bartlett et al. (2008)	Statistic features of 20 AUs in each video segment	Pain	1, 2, 4-7, 9, 10, 12, 14, 15, 17, 18, 20, 23-26	Nonlinear SVM	Self-collected	72.00%
Schmidt et al. (2009)	Maximum speed and amplitude of movement onset of lip corner and eyebrow; AFIA to measure movement	Smile	6, 12, 14, 15, 17, 23, 24, 50	/	Self-collected	/
Saxen et al. (2017)	statistic features (440-dimensional) from the intensity time series of 7 facial AUs	Multiple (6)	1, 2, 4, 6, 9, 12, 25	Rank SVMs	SASE-FE	73.00%
Racoviţeanu et al. (2019)	AlexNet CNN architecture on 12 AU intensities to obtain the features in a tranfer learning manner	Multiple (6)	1, 2, 4-6, 9, 12, 15, 17, 20, 25, 26	SVM	DISFA, SPOS	72.10%

Fig. 3: Overview of muscle movement based SVP detection methods found in [11]



Reference	Method (features)	Expression	Classification	Database	Accuracy
Zhang et al. (2011)	SIFT appearance based features and FAP geometric features	Multiple (6)	RBF SVM	USTC-NVIE	79.40%
Li et al. (2017)	Combining sequential geometric features based on facial landmarks and texture features using HOG	Multiple (6)	Sigmoid	SASE-FE	68%
Mandal and Ouarti (2017)	Fusing subtle (micro) changes by tracking a series of facial fiducial markers with local and gobal motion based on dense optical flow	Smile	SVM	UvA-NEMO	74.68%
Kulkarni et al. (2018)	Combining learned static CNN representations from still images with facial landmark trajectories	Multiple (6)	Linear SVM	SASE-FE	70.20%
Saito et al. (2020)	Combining hardware (16 sensors embedded with the smart eyewear) with software-based method to get geometric and temporal features	Smile	Linear SVM	Self-collected	94.60%

Fig. 4: Overview of hybrid methods for SVP detection found in [11]

(see Table 1) and has concluded that for emotion extraction, lighter and simpler models were more useful.

Model	Reference	Database	Accuracy
Random Forest	Munasinghe [12]	CK+	90%
SVM	Akram et al. [22]	CK+	72% to 100%
Local Threshold Binary Pattern (LTBP)	Shu et al. [23]	BU-3DFE	90%
SVM + WPCA	Zhiguo et al. [24]	BU-3DFE	88.25%
Deep RNN	Yong et. al [25]	BU-3DFE	66.7%

Table 1: Comparison of models found in [12]

An improvement might be to use a *light* neural network for the classifier. One of our metrics of evaluation is the ease of implementation. Kuo et al.'s approach to a robust but compact model in [14] has great performances while being small in size (see fig 5). In order to exploit temporal information (close-up scene extracted thanks to the scene detector) we will have to develop a frame-to-sequence approach which uses image frames as input and then produces a prediction from the whole sequence. In order to use such an approach, we need to implement, as explained in [14],a recurrent neural network into our pre-trained neural network. The solution found in [14] was to add a Gated Recurrent Unit [20] into the frame-to-sequence model(see fig 7).

Method	Ang.	Dis.	Fea.	Hap.	Sad.	Sur.	Neu.	Ave.	Std.	Model size	Ratio
VGG + mSVM[32]	68.52	27.50	35.13	85.32	64.85	66.32	59.88	58.22	18.63	54458K	20.3
AlexNet+mSVM[32]	58.64	21.87	39.19	86.16	60.88	62.31	60.15	55.60	18.68	43501K	16.2
DLP-CNN+mSVM[32]	71.60	52.15	**62.16**	**92.83**	**80.13**	81.16	**80.29**	**74.20**	**12.56**	19655K	7.35
Ours-frame	**82.07**	44.59	41.25	81.01	44.14	**90.12**	75.44	65.52	19.64	**2673K**	**1**
Ours-frame*	74.47	**67.57**	46.88	82.28	57.95	84.57	59.12	67.55	12.78	**2673K**	**1**

Fig. 5: Expression recognition accuracies and the CNN model sizes of different methods on the RAF [16] database found in [14]

The proposed architecture found in Kuo et al.'s paper [14] is a simple convolution neural network (see fig 6), consisting of two convolutions equipped with ReLu and pooling blocks, followed by two fully-connected layers with dropout (to prevent over fitting).

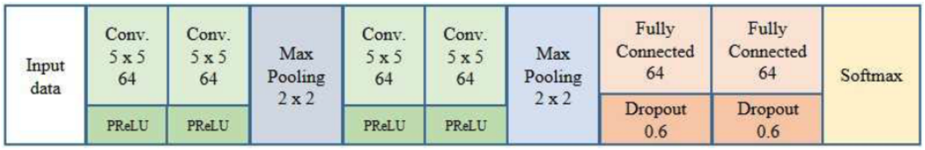


Fig. 6: CNN architecture found in [14]

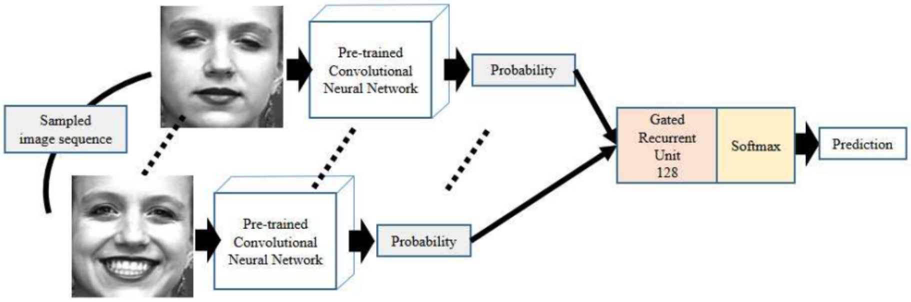


Fig. 7: Framework of the proposed frame-to-sequence approach. The frame-to-sequence model takes features extracted by the pre-trained CNN model and uses their softmax outputs for classification found in [14]

The purpose of the frame-to-sequence model is to facilitate the training process. We would train our model on still labeled images. then, in order to use temporal information, a recurrent neural network such as a LSTM [21] or a GRU [20] to "remember" past information.

Player reactions. Celebrations are exciting moments in a soccer match and an important cue to determine if a segment is an interesting moment of a game. In [15] they train an action recognizer to detect a player celebrating. Two strategies are used to reduce the cost of training data collection and annotation for action recognition. First, using our audio model with a low threshold we could extract segments that are deemed exciting. Second, using still images which are much easier to annotate and and train (less computationally demanding).

4 Extracting metadata for semantic analysis

In this section, we will call metadata any data describing the soccer game that is not in the audio or video streams. This includes the match statistics, but also every single data that can be found that brings some information about the analysed/to-be-processed game. For the purpose of this article we will only focus on data that can be directly mined and/or used by our solution without involving a lot of pre-processing. Furthermore, we will mainly work on the exploitation of statistics and social media interactions as they are two great sources of information that are not often utilized in today's highlight generators.

Classifying footage according to statistical cues. In [18], the authors use what they call "contextual cues". These cues that not only include audio and video analysis, but also imply information such as the basket type (dunk, layup, jumper, 3 pointer, . . .), play-by-play statistics or player ranking (personal statistics for the analysed game and/or season statistics). And this solution can be generalised for nearly all existing sports with some tuning to adapt it to the codes of the analysed sport. We will now adapt this model to soccer games.

First of all, to be able to use play-by-play statistics in the analysis, Bettadapura, Pantofaru and Essa started by aligning them with the videos. To achieve that, they used an Optical Character Recognition (OCR) technique focusing on the "graphic overlay" displaying the scores, the game period and the game clock. They then parse the play-by-play statistics and map the information thanks to the score evolution detected by OCR. *"Example: Say for a particular game, for a particular frame of the video, using OCR on the graphics overlay, we know that the home team score changed from "35" to "38" while the visiting team score was "29" during the "1st half" of the game at game clock "12:22" and when the video timestamp was "28:34". While parsing the play-by-play stats, we see an entry "Player: Jahlil Okafor, Basket Type: 3-Pt Jump Shot, Game Period: 1st Half, Home Score: 38, Visiting Score: 29, Game Clock: 12:22". By matching this entry with the OCR data, we can see that this particular 3-Pt Jump Shot basket by Jahlil Okafor took place at time-stamp "28:34" in the video. This allows us to align the rich contextual info from the stats with the corresponding basket within* *the video."* The same OCR technique can be used for soccer games, along with the available match metadata. We can then use this new information to compute player ranking, score evolution and action type scores.

A quick analysis of existing highlight reels show that moves by star players usually generate more excitement among fans than ones by lower ranked players. The player ranking score will therefore translate this phenomenon in the final result. This score can either be based on the statistics of each player in the analysed game, or their overall statistics of the season. To have the closest score to the ground-truth, it is best to use meaningful season statistics such as the average number of goals/assists per game.

Another cue can be the score evolution. As we know, two games with the same final scoreline

Finally, assuming we can find footage annotated with rich metadata such that each move (skill, long range shot, header, volley, penalty, . . .) is given, we can compute an additional contextual cue for each sequence (especially goals/goal attempts) like it has been done for basketball. To do so, we first have to rank each move according to its excitement level for the viewer. For example an acrobatic goal is considered to be one of the most exciting plays and will most probably be shown in the highlights reel. After these types of moves are ranked to match the user-generated ground-truth as best as it could, each sequence receives a score according to its content with respect to the moves ranking.

These cues are then used in a final computation (normalised weighted average) to get the final score of each footage between 0 and 1, which then allows the model to classify them from most to least emotionally interesting.

Analysis of social media interactions. In the same paper, Bettadapura et al. mention *crowdsourced event summarization*, that is, using social media updates to summarize sport events. Nichols et al. give a detailed presentation and analysis of their affective-based algorithm in their paper [19]. They aim to generate a textual summary of sporting events thanks to status updates on Twitter. Indeed, mining twitter data for relevant tweets and detecting spikes in the volume of tweets can help us identify key moments of the game that the crowd believes is interesting. To be able to generate a coherent text, they construct phrase graphs for each important moment of the event using the corresponding set of tweets, then generate a sentence to summarize the moment.

This method of footage classification can be very useful to generate relevant and emotionally rich highlights. It can also be used to bring additional information to a traditional highlight generation system relying on audio and/or video analysis allowing it to feature referee or fans actions which are not always shown in today's mainstream highlights for example. This type of model can also lead to generating alternative highlights or "lowlights" if social media users express frustration or anger in their tweets. It also gives a lot of importance to the emotions of the supporters/viewers that are not in the stadium. Which could be quite useful nowadays as the COVID-19 pandemic and the lack of supporters in the stadiums for an indefinite period can bias the results given by solutions relying on big variations in audio volume/energy for example.

Pros and cons. In the end, we see that these metadata analysis models can bring a lot of diversity to a generator. The results are therefore a little less similar to what can be seen with ground-truth reels or with traditional algorithms, but they ensure that their content is very rich in emotions. On the other hand, it is clear that using these methods are very costly compared to usual audio and/or video analysis methods. It also requires a more important pre-processing step, which can significantly weigh our solution down.

5 Synthesis

We presented numerous approaches for automatically extracting highlights in soccer games using video and audio streams based on key and emotional moments. We also looked at different tested models and how accurately they extract segments. Our next step is to implement the best of the best and to demonstrate it on soccer games as well as video games.

References

1. Ziyou Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang. “Audio Events Detection Based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework.” In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03)., 5:V–632, 2003. https://doi.org/10.1109/ICASSP.2003.1200049.
2. Mustaqeem, and Soonil Kwon. “A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition.” Sensors 20, no. 1 (January 2020): 183. https://doi.org/10.3390/s20010183.
3. Baijal, A., Jaeyoun Cho, Woojung Lee, and Byeong-Seob Ko. “Sports Highlights Generation Based on Acoustic Events Detection: A Rugby Case Study.” In 2015 IEEE International Conference on Consumer Electronics (ICCE), 20–23, 2015. https://doi.org/10.1109/ICCE.2015.7066303.
4. Radhakrishan, R., Ziyou Xiong, A. Divakaxan, and Y. Ishikawa. “Generation of Sports Highlights Using a Combination of Supervised & Unsupervised Learning in Audio Domain.” In Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, 2:935–39. Singapore: IEEE, 2003. https://doi.org/10.1109/ICICS.2003.1292595.
5. Zhang, Bin, Weibei Dou, and Liming Chen. “Combining Short and Long Term Audio Features for TV Sports Highlight Detection.” In Advances in Information Retrieval, edited by Mounia Lalmas, Andy MacFarlane, Stefan Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, 472–75. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006. https://doi.org/10.1007/11735106_44.
6. Shadiev, Rustam Hwang, Wu-Yuin Chen, Nian-Shing Huang, Yueh-Min. (2014). Review of Speech-to-Text Recognition Technology for Enhancing Learning. Educational Technology Society. 17. 65–84.

7. Sun, X., Q. Yang, S. Liu, and X. Yuan. “Improving Low-Resource Speech Recognition Based on Improved NN-HMM Structures.” IEEE Access 8 (2020): 73005–14. https://doi.org/10.1109/ACCESS.2020.2988365.
8. Rehman, Amjad, and Tanzila Saba. “Features Extraction for Soccer Video Semantic Analysis: Current Achievements and Remaining Issues,” n.d., 11.
9. Xu, Guangyou, Linmi Tao, and Guoying Jin. “SLOW-MOTION REPLAY DETECTION IN SOCCER VIDEOS BASED ON MULTI-LEVEL HMM INTEGRATED WITH SHOT DETECTION,” n.d., 5.
10. Shukla, Pushkar, Hemant Sadana, Apaar Bansal, Deepak Verma, Carlos Elmadjian, Balasubramanian Raman, and Matthew Turk. “Automatic Cricket Highlight Generation Using Event-Driven and Excitement-Based Features,” n.d., 9.
11. Jia, Shan, Shuo Wang, Chuanbo Hu, Paula Webster, and Xin Li. “Detection of Genuine and Posed Facial Expressions of Emotion: A Review.” ArXiv:2008.11353 [Cs], August 25, 2020. http://arxiv.org/abs/2008.11353.
12. Munasinghe, M. I. N. P. “Facial Expression Recognition Using Facial Landmarks and Random Forest Classifier.” In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 423–27. Singapore: IEEE, 2018. https://doi.org/10.1109/ICIS.2018.8466510.
13. V. Kazemi and J. Sullivan, ”One Millisecond Face Alignment with an Ensemble of Regression Trees,” in Computer Vision and Pattern Recognition (CVPR), Columbus, 2014.
14. Kuo, Chieh-Ming, Shang-Hong Lai, Michel Sarkis, and San Diego. “A Compact Deep Learning Model for Robust Facial Expression Recognition,” n.d., 9.
15. Merler, Michele, Dhiraj Joshi, Quoc-Bao Nguyen, Stephen Hammer, John Kent, John R. Smith, and Rogerio S. Feris. “Automatic Curation of Golf Highlights Using Multimodal Excitement Features.” ArXiv:1707.07075 [Cs], July 21, 2017. http://arxiv.org/abs/1707.07075.
16. Li, Shan, Weihong Deng, and JunPing Du. “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild.” In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2584–93. Honolulu, HI: IEEE, 2017. https://doi.org/10.1109/CVPR.2017.277.
17. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, ”The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.
18. Pantofaru, Caroline, Irfan Essa, and Vinay Bettadapura. “Leveraging Contextual Cues for Generating Basketball Highlights.” In Proceedings of the 2016 ACM on Multimedia Conference - MM ’16, 908–17. Amsterdam, The Netherlands: ACM Press, 2016. https://doi.org/10.1145/2964284.2964286.
19. Nichols, Jeffrey, Jalal Mahmud, and Clemens Drews. “Summarizing Sporting Events Using Twitter.” In Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces - IUI ’12, 189. Lisbon, Portugal: ACM Press, 2012. https://doi.org/10.1145/2166966.2166999.
20. Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.” ArXiv:1412.3555 [Cs], December 11, 2014. http://arxiv.org/abs/1412.3555.
21. Staudemeyer, Ralf C., and Eric Rothstein Morris. “Understanding LSTM – a Tutorial into Long Short-Term Memory Recurrent Neural Networks.” ArXiv:1909.09586 [Cs], September 12, 2019. http://arxiv.org/abs/1909.09586.

22. A. Alsubari, D. N. Satange and R. J. Ramteke, ”Facial expression recognition using wavelet transform and local binary pattern,” in 2nd International Conference for Convergence in Technology (I2CT), Mumbai, 2017.
23. S. An and Q. Ruan, ”3D facial expression recognition algorithm using local threshold binary pattern and histogram of oriented gradient,” in IEEE 13th International Conference on Signal Processing (ICSP), Chengdu, 2016.
24. Z. Niu and X. Qiu, ”Facial expression recognition based on weighted principal component analysis and support vector machines,” in 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE), Chengdu, 2010.
25. Y.-G. Kim and X.-P. Huynh, ”Discrimination Between Genuine Versus Fake Emotion Using Long-Short Term Memory with Parametric Bias and Facial Landmarks,” in IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, 2018.