Highlight Detection with Audio and Visual Analysis for Soccer Matches Summarization

Ben Ohayon [*] [1] Knaan Koosh [*] [1]

Abstract

Nowadays sport events are highly covered through many mediums. To be able to cover as many events as possible, each event needs to be summarized in a short review – usually a highlights videos. Each highlight video requires a significant amount of man hours of picking the right scenes, editing, etc. These reviews needs to contain all the meaningful events from the game – goals, attempts to score, major fouls, rare events (a passionate fan running naked in the field) and more. Our main goal is to create a model which classifies a video segment from a sport event as a highlight or a regular one. We limit ourselves in this project to soccer matches by a specific dataset, but we believe that any sport with a similar "flow" to soccer can be classified by our model with the right dataset.

1. Introduction

1.1. Motivation

Soccer is a well-known sport which attracts many viewers and fans. The great majority of the games are watched live primarily by the fan base of the playing teams, other soccer fans are kept updated by articles, reviews and recaps. Reviews for these games are essential and require a lot of work to be made. Another problem about manually editing these games is bias inserted by human hand about the teams, players, referees and more. An automated procedure that can classify every moment as highlight or not with high precision can solve these issues easily. We suggest a novel method to automate the process of deciding whether a short video is considered a highlight or not.

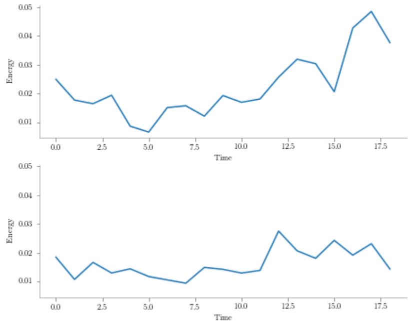

Figure 1. Two energy signals derived from the audio signal. The upper is of a highlight and the lower is of a non-highlight.

1.2. Problem

Highlight videos and regular videos are relatively similar and differentiating between them may be hard, furthermore many types of scenes may be a highlight – controversial referee decisions, goals and free kicks are all an example of moments of high interest but there is a great difference between the scenes in both the visual and auditive aspect. Our problem is to find and learn scene characteristics descriptive enough to differentiate between a highlight and a regular scene (For example, in Figure. 1 we can see the volume of an audio signal for a highlight and a regular scene) and we believe a combination of visual and auditory features might work.

1.3. Approach

The model needs to be general enough to account for many aspects of the game even if they are variant in nature (in a visual or auditory sense). Our idea is to consider video and audio streams and make a model which uses both streams to decide if the video sequence is a highlight. Our prior research revealed that there is a high correlation between highlights and the audio signal in terms of volume. We use 3D Convolutional network to extract spatiotemporal features from the videos and a LSTM network to analyze the audio stream and combine their outputs to get an accurate prediction based on all the information provided.

[*] Equal contribution [1] Tel Aviv University, Tel Aviv, Israel. Correspondence to: Ben Ohayon <benohayon@mail.tau.ac.il>, Knaan Koosh <knaankoosh@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

2. Related Work

As defined above, a highlight can contain a few different events, whether a goal, foul or a substitution, a highlight is still a specific moment of special interest. Classifying a segment is thus a difficult task since finding a common denominator in clips that are visually different might prove hard. In our review we put emphasis on not only Highlight Detection but also on the components that made the classification possible such as Action Recognition, Audio Analysis and LSTM Networks.

2.1. Action Recognition

Action recognition in videos has been a task that attracted many researchers. Since the release of large scale classified video datasets such as Google's Youtube-8M Dataset (Abu-El-Haija et al., 2016), University of Florida Central's UCF101 (Soomro et al., 2012) and more, researchers have developed interesting techniques and network architectures to classify the action being done in the videos. Due to the continuous nature of videos there was a need to extract temporal features from the data, one of few approaches (Tran et al., 2014) proposed to use 3-Dimensional Convolutional Networks to learn spatiotemporal features, they achieved impressive results on the UCF101 dataset and furthermore their architecture is pretty simple which led us to use it. Another interesting take was (Simonyan & Zisserman, 2014) which inspired us to build a two-stream structure, Karen & Andrew proposed an architecture that consists of two ConvNets streams in which one takes single frames in order to extract spatial features and the other takes an optical flow matrix calculated from a few consecutive frames in order to extract temporal features.

2.2. Highlight Detection

Highlight detection in sports has been a subject to a lot of research lately. As the global popularity of soccer continues to grow, companies are accumulating large amounts of data and metadata and researchers utilize it to create novel techniques to tackle our problem. Due to the diverse nature of highlights, (Godi et al., 2017) proposed to analyze footage of hockey audience using 3D-ConvNets in order to determine which moment is considered exciting. Another approach (Baillie & Jose, 2004) built an elaborate dataset that consists of multiple types of audio segments during a game such as speeches, crowd chanting or cheering, etc. and used HMM to classify the segments, ones classified as cheering were considered key events.

2.3. Automatic Summarization

Although the primary focus of this project is on sports videos, attempts to summarize first person videos such as phone camera videos have contributed great insight to our final goal which is automating the creation process of highlight videos, (Yao et al., 2016) proposed a ranking system that consists of a spatial and temporal streams, the first is a pretrained AlexNet (Krizhevsky et al., 2017) and the second is a 3D ConvNet, both their outputs are combined with a dense layer ultimately outputting a rank. By ranking each segment of the video a rank-time series is made and a summarization can be created by taking time areas around a selected amount of peaks along the series.

3. Our Approach

In this section, we propose a new architecture that consists of an audio and video streams combined at their end with a fusion layer. The idea is to leverage both the audio and visual characteristics of a highlight to create a feature descriptive enough to be able to differentiate between regular clips and a wide variety of highlights. These features will eventually be run through fully connected layers in order to get a final prediction. The simple illustration of the model's architecture can be seen in Figure 2.

To optimize our model we used an Adam optimizer (Kingma & Ba, 2014) which is a method for stochastic optimization with adaptive learning rate. Our initial learning rate was set to $1\mathrm{e}{-5}$ with a batch of size 5 due to memory limitations.

3.1. Audio Stream

As said earlier, we found a high correlation between a highlight and its audio signal characteristics such as sudden peaks and energy. The Audio Stream starts with a preprocessing step, we sample the video's audio signal X at 22KHz and produce an energy signal $\boldsymbol{E}$ by a simple RMS calculation:

$$\boldsymbol{E}[\boldsymbol{t}] = \sqrt{\frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\boldsymbol{0}}^{\boldsymbol{n}-\boldsymbol{1}} \boldsymbol{X}[\boldsymbol{i}]^{\boldsymbol{2}}}$$

This yields us a signal the length of the video and is practically the audio stream's input. The audio network is a simple implementation of 2 layers of bi-directional LSTM (Hochreiter & Schmidhuber, 1997) with 10 hidden states. The intent here is to learn sudden changes in volume through time that might be indicative of exciting moments. Eventually the values of the hidden states of the last layer (which we consider to be the audio features) are passed on to the fusion layer to be combined with the spatiotemporal features.

3.2. Video Stream

Videos naturally, can be decomposed into spatial and temporal components. As (Tran et al., 2014) suggested, we can learn spatiotemporal features from videos by using 3D ConvNets. 3D ConvNets use a three dimensional kernel

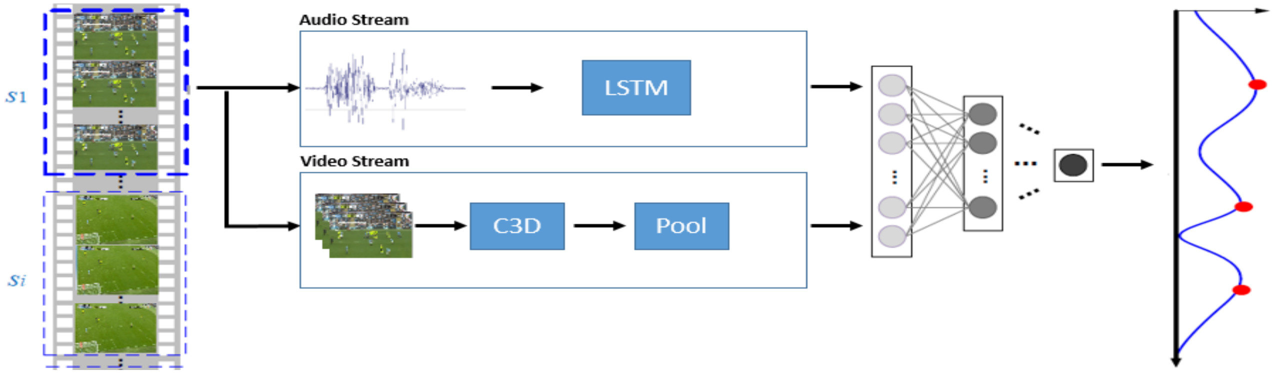

Figure 2. A Highlight Detection model illustration. The input video splits into two streams - audio and video. The audio signal is analyzed by a LSTM network. The video stream is analyzed by a C3D network. The output of the two streams is combined and a fully connected layer outputs a ranking curve.

Table 1. Details of the C3D network architecture, c3d, pool3d and fc refer to 3D Convolution, 3D Max Pooling and Fully Connected layers.

	c_1	p_1	c_2	p_2	c_{3a}	c_{3b}	p_3	c_{4a}	c_{4b}	p_4	c_{5a}	c_{5b}	p_5	fc_6	fc_7
type	c3d	pool3d	c3d	pool3d	c3d	c3d	pool3d	c3d	c3d	pool3d	c3d	c3d	pool3d	fc	fc
size	3	64	64	128	128	256	256	256	512	512	512	512	512	512	4096
channels	64	64	128	128	256	256	256	512	512	512	512	512	512	4096	30

and convolve it with a stack of consecutive frames (video) and thus preserving temporal features. The network input is stack of downsampled frames (256x256 input resolution), usually a stack consists of 15 evenly distributed and consecutive frames. We use a similar architecture to the proposed C3D in (Tran et al., 2014), the network consists of 8 convolution layers, 5 pooling layers and 2 fully connected layers. Each of the convolution layer has a kernel size of 3x3x3 with a stride of 1x1x1, the pooling layers have a kernel size of 2x2x2 with a stride of 2x2x2 (with the exception of the first pool which has kernel 1x2x2 and stride 1x2x2), lastly the fully connected layer have output of 4096 units. Table 1. summarizes the video stream layers.

3.3. Ranking

Ultimately to make a classification we calculate a probability by using the outputs of our two streams, namely the audio and spatiotemporal features. The two sets of features are concatenated and passed through three fully connected layers with ReLU activation function, then we use a softmax layer to calculate a probabilty of the clip being a highlight and call it a rank. To create a summarization of a match we segment it to random length (10-30 seconds) clips and run them through the model to calculate a rank for each. We pick around fifteen peaks with the highest rank and concatenate clips from the time area around the peaks to create a highlight video of fifteen highlights.

4. Dataset

Although there is a great interest in Action Recognition and Highlight Detection for sports videos, public datasets were too small or did not fit our case, for example Youtube8M, Sports1M or UCF101 contained videos of various sports or in specific spaces like living rooms, kitchens, etc. in either case we were unable to find a fitting large scale dataset due to the specific nature of our goal and copyright limitations. We evaluate our model on a newly created dataset including around 8 hours of match play. The dataset is made of 8 full soccer matches played by different teams in the last year, the matches were cut into 10 to 30 seconds long videos and annotated according to the matching official highlights video. The dataset in its final form contains more than 2000 video samples split into highlight and non highlight clips to learn from.

5. Experiments

Since previous work did not include the use of both visual and audio signals, our main purpose was to check the influence of each stream on the classification accuracy. To do so, we have conducted three different experiments. The first would be for the video stream alone, the second for audio stream and the third for our main model that combines the two streams. Finally we intend to create an highlight video using the model that yielded the highest results.

5.1. Data

All experiments were tested on a part of our newly created dataset that contains 500 highlight clips and 700 non highlight clips, the dataset was split into training and validation sets with a ratio of 80%-20%.

5.2. Training Parameters

In the experiments we use a sample rate of 22KHz for the audio signal and we sample 15 consecutive and evenly distributed frames from the video clip. For our optimizer we chose a learning rate of $\mathbf{1e{-}5}$ and use a batch size of 5 for all experiments with the exception of the audio stream which used a batch size of 20. The models were trained for 50 epochs due to the fact that around the 40th epoch all models were showing signs of overfitting.

5.3. Evaluation Metrics

The ideal way of evaluating our model is to let a human subject decide how exciting our automatically generated highlight videos are and comparing them to human made match recaps. Since evaluating in the ideal way might take a long time we evaluate the model by calculating the average precision and recall of highlight detection in the validation set for 5 runs. Finally we present the classification accuracy for the videos in the validation set.

5.4. Compared Approaches

We compare each of the experiments results with approaches that use visual or audio based techniques. We chose ones that achieved impressive results and use a similar way of evaluation for easier comparison. The approaches:

1. Indirect Highlight Detection (Godi et al., 2017). This approach uses audience footage to determine exciting moments in hockey matches.
2. Cricket Highlight Detection (Tang et al., 2011). This approach uses cricket matches footage with HMM and LSVM to classify video segments.
3. Audio Based Event Detection (Baillie & Jose, 2004). an audio based approach that uses HMM as a classifier
4. Highlight Extraction for TV Baseball Programs (Rui et al., 2000). Another audio based approach that uses various classification techniques such as SVM and KNNs.

5.5. Results

We present the results carried out to evaluate our approach to highlight detection. The first experiment measured the accuracy, precision and recall of the audio stream, the second of the video stream and the third of the combined model. The acronyms presented in Table. 2 are ordered in the same order as the Compared Approaches section and represent them accordingly.

Table 2. Result comparison between the experiments and the approaches mentioned above.

	Audio	Video	Combined	IHD	CHD	ABD	HBP
Prec.	0.78	0.65	0.88	0.69		0.70	
Rec.	0.71	0.60	0.68	0.84		0.90	
Acc.	0.78	0.63	0.83	0.78	0.88		0.75

We have shown good results comparing to previous work with the same goals as can be shown in Table. 2. As we assumed from the beginning, the audio signature of one moment in a soccer game can tell a lot about the scene that is to say how much interest exists in that scene which lead to higher classification accuracy with audio stream. The video stream was relatively hard to train due to the visual similarity between a highlight and non highlight clips and took much more computation time, still the results were a bit disappointing. At last, we combine the two streams to build our model and got even better results compared to the audio stream and even relatively good compared to other approaches.

6. Conclusions and Discussion

We have developed a two-stream audio/video model for highlight detection in soccer matches in this paper. The model extracts features from both spatial and temporal dimension by a 3D convolution and analyzes audio features using LSTM cells. Each stream generates multiple channels of information and final feature representation is processed through a fully connected layer. As shown in the final results we see an increase of accuracy for our model compared to both streams separately. A Possible reason for this increase is a compensation from the audio signal for visual characteristics such as similarity between positive and negative highlights or a difference in color and viewing angles in the highlight clips due to the different stadiums, teams and audience shown in the videos. As said in prior papers on the same subject, maybe more can be achieved with deep belief networks which achieves promising performance on object recognition tasks, another suggestion would be to use gray scale frames in order to achieve higher uniformity between different matches in the price of visual information loss.

7. Software and Data

Our code and model can be found in the following link: https://github.com/knaankoosh/HighlightSoccerNet
An automated summary video be uploaded soon.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016. URL http://arxiv.org/abs/1609.08675.

Baillie, M. and Jose, J. M. An audio-based sports video segmentation and event detection algorithm. pp. 110–110, June 2004. doi: 10.1109/CVPR.2004.298.

Godi, M., Rota, P., and Setti, F. Indirect match highlights detection with deep convolutional neural networks. *CoRR*, abs/1710.00568, 2017. URL http://arxiv.org/abs/1710.00568.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. 9:1735–80, 12 1997.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL http://doi.acm.org/10.1145/3065386.

Rui, Y., Gupta, A., and Acero, A. Automatically extracting highlights for tv baseball programs. pp. 105–115, 2000. doi: 10.1145/354384.354443. URL http://doi.acm.org/10.1145/354384.354443.

Simonyan, K. and Zisserman, A. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014. URL http://arxiv.org/abs/1406.2199.

Soomro, K., Zamir, A. R., and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

Tang, H., Kwatra, V., Sargin, M. E., and Gargi, U. Detecting highlights in sports videos: Cricket as a test case. *2011 IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2011.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., and Paluri, M. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014. URL http://arxiv.org/abs/1412.0767.

Yao, T., Mei, T., and Rui, Y. Highlight detection with pairwise deep ranking for first-person video summarization. pp. 982–990, June 2016. ISSN 1063-6919. doi: 10.1109/CVPR.2016.112.