Automated Sports Highlight Detection: Based on Audience Reactions

Sanskruti Shah
College of Information and Computer Science
UMass Amherst
sanskrutirah@cs.umass.edu

Shasvat Desai
College of Information and Computer Science
UMass Amherst
shasvatmukes@cs.umass.edu

Debasmita Ghose
College of Information and Computer Science
UMass Amherst
dghose@cs.umass.edu

Abstract

Highlights in a sports video are the key exciting moments in the match which attract attention of the spectators in the match. A considerable amount effort is spent in extracting such highlights from the match requiring a lot of investment in terms of time and cost where the domain experts decide which frames must be included in the highlight, thus making it an expensive process, so there should be ways of generating automated highlights. In order to do so, we experiment with different methods to generate automatic highlights, using 3D Convolutional networks, HOG-SVM and pre-trained models on similar sports datasets. A popular method is to train a classifier on the features extracted by looking at the game play. Instead of that, here we reverse the paradigm by performing action recognition on spectator reactions to the game, which enables us to identify the highlights in the game. We use audience reactions to classify frames in the game as "highlights" and "standard play", because this method can be generalized to detect highlights in any sport. For this purpose, we have used the S-HOCK dataset [8], [9] which contains videos of spectators watching an ice-hockey match. We obtained an accuracy of 26% using HOG-SVM, 67% using pre-trained models and 74% using 3D-CNN.

1. Introduction

Sport highlights generation, is the task of creating a summary of a sport event that gives the viewer a summary of the game through its key moments. According to [1], highlights are those video segments that are expected to excite the users the most. In this paper we attempt to solve this problem by automatically learning visual features using deep architectures, using handcrafted features with linear classifier and using transfer learning models to discriminate between highlights and ordinary actions.

Figure 1. Example video sequences of a goal event (left) and standard play time (right).[1]

In general, sports highlights are created manually by experts, and therefore it is a labor-intensive task. Due to the increasingly large amount of videos of the sports events happening everyday, there is an increase in the demand for automatic tools for highlight generation of sports video. Our work attempts to solve this problem by analyzing the audience behavior to identify changes in actions, that could only be caused by a highlight in the game, like clapping, standing and waving.

We experiment with multiple techniques like HOG-SVM, 3D-Convolutional Neural Networks and using transfer learning on pre-trained networks trained on various action-recognition datasets.

The rest of the paper is organized as follows: in Section 2 we summarize similar work that have attempted to generate sports highlights, using various other techniques, in Section 3 we describe the dataset used and the preprocessing that was performed on the data. Section 4 talks about the different models that were used to solve the problem, while Section 5 we compare the results obtained by the different models. Lastly, in Section 6 we draw some conclusions and describe the scope for future work.

2. Literature Review

There has been considerable amount of work that has been done in the domain of sports video summarization and the work by Shih, Huang-Chia [2] has shown a very detailed literature survey of all these methods, categorizing the methods on the basis of the type of sport and the type of method used for action recognition.

Most of the work in highlight generation from sports videos is done by analyzing the actions of the players and tracking the objects of interest throughout the game-play ([3], [4], [5]).

Our work reverses the paradigm in recognizing important events in the game, as shown in the work of Godi et.al. [1] and Conigliaro et. al [2]. The work [1] uses a dense 3D - Convolutional Neural Network with a fixed spatial dimension of 100 x 100 and the temporal dimension of 30 frames. Their network consisting of two convolutional layers with 12 filters of size 3x3x3 each, followed by two max pooling layers of size 2x2x2 followed by two convolutional layers with 8 filters each of size 3x3x3 and three fully connected layers of decreasing dimensionality (containing 32, 8 and 2 neurons respectively), with ReLU activation in each convolutional layer, takes an input as a spatial temporal cuboid of size 100 x 100 x 30. Their network achieves the task of classifying the frames into ”Highlights” and ”Standard Play”.

3. Dataset

The dataset used by us is called the S-HOCK dataset [8], [9], made public by the Vision and Image Processing Lab of University of Verona, Italy. The dataset contains videos of spectators from multiple cameras, watching Ice Hockey matches at Trento (Italy) during the 26th Winter Universiade and the corresponding game-play for those time frames. Those videos have been annotated to label any activity among the spectator crowd when there is an important event in the game and the cause of the activity (i.e. score, foul etc., with the team which scored). The videos were captured with the resolution of 1280x1024 at 30fps. From each match, they selected a pool of sequences which were representative of all possible situations, to be a part of the dataset.

The dataset contains multiple annotated videos of duration 30 seconds. We extracted frames at 30fps, which gave us about 900 frames per video. We resized the images from the original resolution to 256x256 in order to increase the computational efficiency without any significant loss in information. We also tested our models by augmenting the frames by increasing the number of positive samples in the dataset, in order to tackle the problem of data imbalance, since the number of negative samples were significantly higher than the number of positive samples. We also stitched multiple videos belonging to the same match, extracted frames at 30fps, resized them and tested our models.

4. Model

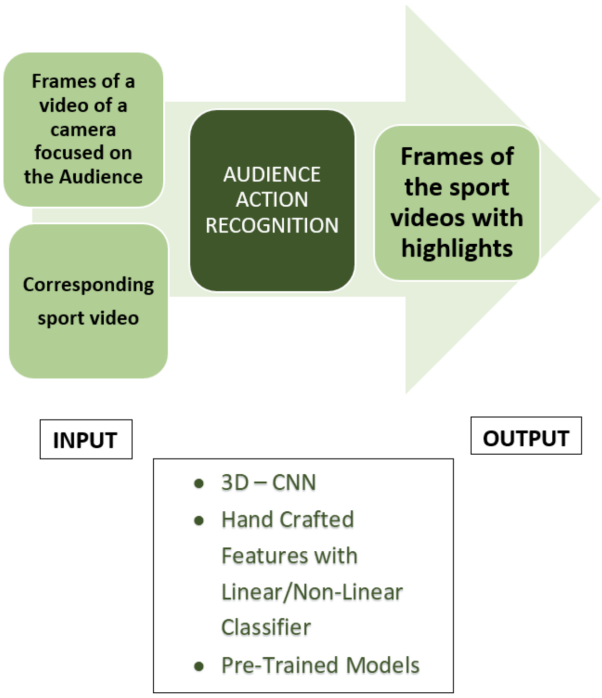

Figure 2. Overview of the project

4.1. 3D Convolutional Neural Networks

We used a 3D convolutional network which had spatial dimension as its first two dimensions and the third dimension as the time. Hence our input to the network was a spatio-temporal cuboid with each cuboid capturing a 1 second of the match. Since we sampled the video at 30fps, the temporal dimension of our input cuboid is 30.

4.1.1 Network Architecture

The architecture of the network is as shown in Figure 3. The network takes as input video cuboids of 256x256x30, where

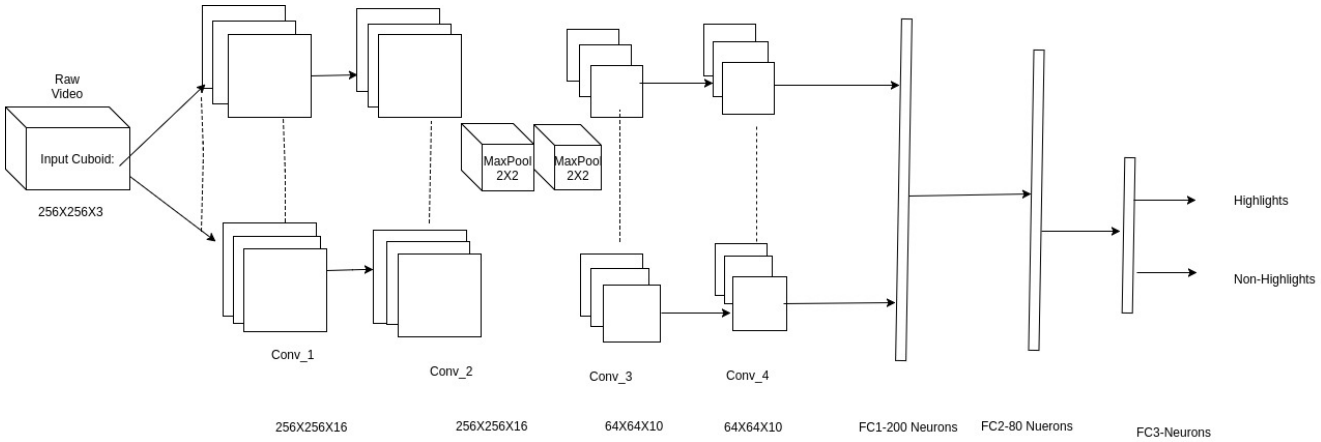

Figure 3. Network Architecture

the first two numbers refer to the spatial dimension while the third is the temporal depth (number of frames). The first two convolutional layers are composed 16 filters 3x3x3, to capture spatio-temporal features from the raw data. These are followed by a 2x2x2 max pooling layer to detect features at different scales. In the latter two convolutional layers, 10 3x3x3 convolutional filters have been used. In all convolutional layers the ReLU activation has been used. The network is then unfolded with a flatten layer followed by three fully connected layers of decreasing dimensionality (200, 80, and 2 neurons respectively). The final classification task is achieved by a softmax layer that outputs the probability of a test sample to belong to each of the two classes: "highlight" and "standard play". The activation used for the convolution layer was ReLU and a dropout of 0.5 was added.

4.1.2 Experiments

We considered the network specified in the work of Godi et. al [1] as a model network for detecting highlights from frames of a video. In order to tune our network, we experimented with the number of filters in the range 10 to 128 to enable it to capture various features and experimented with the number of convolutional layers in the range 1 to 4 and added 2 max-pooling layers. Further we added an additional fully connected layer and varied the number of neurons in both the fully connected layer in the range, 50 to 264. We implemented our models using the Keras/TensorFlow framework on the GPU provided by Amazon Web Services.

4.2. HOG-SVM

We experimented with a handcrafted feature extraction method - Histogram of Oriented Gradients and then applied a linear SVM to classify the frames as "highlights" and "standard play".

The frames were extracted from the video at 3fps, and each frame was resized to (256, 256). We experimented

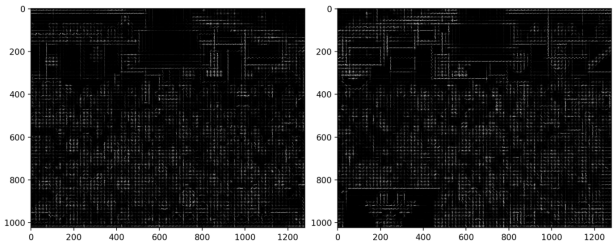

Figure 4. HOG representation - standard play v/s highlight

with different cell sizes and block sizes. The cell sizes tried were 16x16 and 24x24, with block sizes of 4x4 and 8x8. The best results were obtained with 16x16 pixels per cell and 4x4 cells per block, the HOG features were training using a SVM classifier with RBF kernel and C=1.

The accuracy for this model was 25.67% where 77 frames were detected as a highlight from the 300 frames corresponding to 10 seconds of audience excitement in the test video.

4.3. Pre-Trained models

We experimented with pre-trained models trained on HMDB-51 [10] and KTH datasets [11] which are similar to our dataset. We removed the last fully-connected layer (this layers outputs are the 51 class scores for a different task like HDMB-51), then treated the rest of the ConvNet as a fixed feature extractor for the new dataset. Once we extracted the CNN codes for all images, we trained various classifiers such as Linear SVM and Artificial Neural Networks for our dataset. We also loaded this pre-trained model from checkpoints and fine tuned it for our dataset. The resultant accuracy on our dataset was 67% on our dataset.

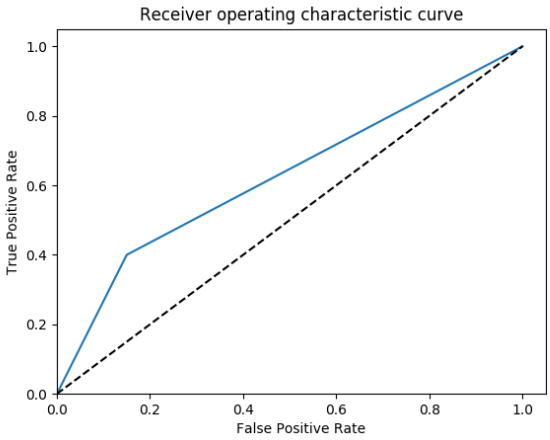

Figure 5. ROC curve for 3D CNN

5. Results

The 3D CNN provided the best results from the above models with an accuracy 74% then the pre-trained model using KTH dataset with an accuracy of 67% and finally the HOG-SVM model with 26%. The ROC Curve for the 3D CNN model is shown in Figure 5. The model might be returning false negatives due to factors like empty seats, truncated image size due to computational power constraints and small size of the dataset.

The detected highlight from a small section of an ice hockey match between Kazakhstan and USA using the 3D CNN model can be seen on `https://youtu.be/y4Wt10KcIe8`.

6. Conclusion

In this project, we implemented a method to temporally locate highlights in a sport event by analyzing the audience behavior. We used a deep 3D Convolutional Neural Network on cuboid video samples to discriminate between different excitement of the spectators and compared it to a handcrafted feature model and used transfer learning to use models trained of different datasets. The 3D Convolutional Neural Network produced a good score which is proportional to the probability of having an interesting highlight in that time location, thus enabling the model to identify goals and other highlight worthy moments.

There is still further scope of improvement. In our opinion, the main limit of this model is in the way we take into account the temporal information; indeed we extend a standard CNN to work with 3D data, where the third dimension is time. Improved models such as recurrent neural networks (RNN) and long-short term memory (LSTM), which capture the time dimension well could enhance the results.

References

[1] Godi, Marco, Paolo Rota, and Francesco Setti. "Indirect Match Highlights Detection with Deep Convolutional Neural Networks." arXiv preprint arXiv:1710.00568 (2017).

[2] Shih, Huang-Chia. "A Survey on Content-aware Video Analysis for Sports." IEEE Transactions on Circuits and Systems for Video Technology (2017).

[3] Tang, Hao, et al. "Detecting highlights in sports videos: Cricket as a test case." Multimedia and Expo (ICME), 2011 IEEE International Conference on. IEEE, 2011.

[4] Assfalg, Jrgen, et al. "Semantic annotation of soccer videos: automatic highlights identification." Computer Vision and Image Understanding 92.2 (2003): 285-305.

[5] Assfalg, Jrgen, et al. "Semantic annotation of soccer videos: automatic highlights identification." Computer Vision and Image Understanding 92.2 (2003): 285-305.

[6] Ekin, Ahmet, A. Murat Tekalp, and Rajiv Mehrotra. "Automatic soccer video analysis and summarization." IEEE Transactions on Image processing 12.7 (2003): 796-807.

[7] Conigliaro, Davide, et al. "ATTENTO: ATTENTion Observed for Automated Spectator Crowd Analysis." HBU. 2013.

[8] Conigliaro, Davide, et al. "The s-hock dataset: Analyzing crowds at the stadium." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[9] Berlonghi, Alexander E. "Understanding and planning for different spectator crowds." Safety Science 18.4 (1995): 239-247.

[10] Kuehne, Hilde, et al. "HMDB51: A large video database for human motion recognition." High Performance Computing in Science and Engineering 12. Springer, Berlin, Heidelberg, 2013. 571-582.

[11] I. Laptev. Local Spatio-Temporal Image Features for Motion Interpretation. PhD thesis Department of Numerical Analysis and Computer Science (NADA) KTH S-100 44 Stockholm Sweden 2004. ISBN 91-7283-793-4.