Journal of Intelligent Learning Systems and Applications, 2022, 14, 107-114
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Video Highlight of Farcana Streaming from Twitch

Ilman Shazhaev[1], Dimitry Mihaylov[2,3], Abdulla Shafeeg[3]

[1]Management Department, Farcana, United Arab Emirates
[2]China Branch of BRICS Institute of Future Networks, Shenzhen, China
[3]Science Department, Farcana, United Arab Emirates
Email: abdulla.shafeeg@farcana.com

How to cite this paper: Shazhaev, I., Mihaylov, D. and Shafeeg, A. (2022) Video Highlight of Farcana Streaming from Twitch. *Journal of Intelligent Learning Systems and Applications*, **14**, 107-114.
https://doi.org/10.4236/jilsa.2022.144009

Received: September 11, 2022
Accepted: November 5, 2022
Published: November 8, 2022

Abstract

The Highlights are the most interesting, selling moments from video stream, which can make the viewer watch the entire video. They are like a shop window: everything that is bright and colorful goes there. Seeing them, the user can understand in advance what is inside the video. And they are more versatile than trailers: they can be made shorter or longer, embedded in different places in the user interface. The user sees a selection of highlights as soon as he gets to the website or watches a video clip on YouTube or even a section of a stream of a popular blogger/influencer. The user's attention is immediately attracted by the most memorable shots. Naturally, it is becoming more tedious to manually create all video highlights, due to the immense amount of material that the highlights are needed for. Thus, creating an algorithm, capable of automating the process would make the process significantly simpler. Besides easing up the work, this process will pave the way to a whole set of new applications that before did not seem real. At the same time, this would be a new process where the AI would not need to be fully supervised by a human, but be capable of identifying and labeling the most interesting and attractive moments on screen. After doing a literature review on video highlight detection, this paper has utilized the model presented in the study by [1] to determine the possibility of attaining highlights from the Farcana 2.0 version Twitch video feed.

Keywords

Highlights, Intelligent System, AI, Algorithms

1. Introduction

One cannot say that creating an automated AI algorithm to decrease the level of

supervision during the process of highlights creation has not been attempted before. On the contrary, numerous scholars have attempted and succeeded to make the highlights gathering process much easier. This paper focuses on the study attempted by [1]. Scholars have made it their priority to develop a way to automate the process of having the AI automatically gather the most interesting moments in on-screen production. Their method of contrastive learning is undeniably complex, yet it has proven to be a success. Other scholars have attempted studies mostly focusing on solutions that required full-time supervision of the process by a human being. Yet, in study [1], they have managed to achieve an automated solution decreasing the amount of supervision necessary. Their framework encoded the video "into a vector representation" [1], thereby teaching the AI to pick the most interesting moments that could be classified as video highlights and further demonstrated to a large audience. Their method was compared to the existing solutions requiring state-of-the-art approaches and equipment heavily relying on substantial input from external data. Yet, the approach in [1] provided us with a simpler solution requiring much fewer resources than say, the LM system which employs as many as 10 million videos in the process of training their algorithm.

This algorithm can be applied in video retrieval, finding recommended content, help in browsing, and naturally editing. Although this particular paper is focused on setting the task of receiving the best highlights of the Farcana streaming on Twitch utilizing the gameplay of the Alpha 2.0 version release. The application of method presented in [1] allows us to update the innovative approach and automate the process of video highlight retrieval. This saves on manpower and time, at the same time saving on costs.

What happens in the frame, whether it's an explosion, a gunshot, or a fatality move, affects the conversion in different ways. This influence can be measured—some statistics allow you to guess in advance which fragments of the film are best suited for the role of highlights [2]. And then special slicing algorithms are turned on part of a large system for preparing content for streaming. It can create a highlight from any video file and at any time after it enters the system. The operator creates a job to generate a highlight, and the system processes it according to the set priorities.

There are many algorithms. We have used several in our experience. The first is the scene detection algorithm. He must determine the beginning and end of the desired piece of the film. In the course of training, the algorithm was first shown various shooting angles, taught to find them, and then, from the selected angles, they were taught to highlight specific scenes. Also, information from the video and audio stream is used to determine suitable segments of the film. This is necessary so that dialogs are not interrupted when creating highlights.

The second is the scene selection algorithm for creating a highlight. We trained our algorithm based on Twitch streams of Farcana gameplay. We took all the streams that were available in the Alpha 2.0 version released in the second quarter of 2022 and available only under Private Sale terms. Now, having a trained

algorithm, we can rank any scene of a new stream file, calculating the probability of how this scene is suitable for creating a highlight.

When the set of scenes is ready, the algorithm selects the final candidates, from which the highlight will turn out. At the same stage, one can do image processing, color correction, and filters. The results are reviewed by a post-moderation person. The latter can always reject a ready-made highlight if it turned out badly. If the highlight passes the test, it is placed on special servers, from where it is delivered to users using a CDN.

2. Literature Review

Video Highlight Detection has gone through a substantial path of development. Of course, initial attempts were made in the sphere of sports with a focus on sporting highlights. Then this trend moved forward to social media and first-person feeds [3]. However, at the start of this path most methods applied were focused solely on the heavily supervised approach having an individual moderate the content and general feed.

Further steps in video summarization have started to become predominantly unsupervised with scholars relying on heuristics and representativeness to be able to obtain a summary [3]. At the same time, weakly supervised methods have been developed for the production of video annotations enabling users to benefit from video tagging. Nonetheless, a substantial number of scholars have attempted to shift progress in R&D on unsupervised models of highlights capture. For example, [4] have acknowledged that AI is at the forefront of technological development and the shift to a new level of global operation. These scholars have noted the immense success of WSC Sport as a sports algorithm framework capable of providing highlights of NBA team statistics, highlights, and various kinds of data to sports fans in real-time. An automagical solution is what WSC Sport calls themselves and explains with the definition of the word from a Dictionary: something happens automatically, but from the outside, no one can explain exactly how. The initially Israeli startup WSC Sport broke into the American market in late 2015 and turned the NBA's highlighting process around [5]. There, video reviews are especially needed, because dozens of effective actions are performed in each game, and a fan needs to highlight the main ones.

WSC Sports Technology is an artificial intelligence that can collect highlights from really key moments without any human intervention at all. The technology recognizes the signs of the most important episodes: how the score changes after an accurate throw, how emotionally the players and fans react, whether the commentator powerfully intones, and so on [4] [5]. During the match, the algorithm generates a rating from all game segments in general, each of which is marked with details: the characters, the type of throw (long-range, medium, from under the ring, with resistance or open), the features of the episode (quick attack, positional attack, throw immediately after transmission) [4] [5]. The technology allows you to create highlights at any time and for any duration, just set these parameters. For example, cuts of the actions of a particular player fit per-

fectly into a live broadcast, if he is especially (or unexpectedly) good, in a long break or immediately after the match, a video of the main moments as a whole is organic. A club or an owner of media rights can use customized highlights at any time and for any reason: if one wants to highlight the main antagonist of a winning match—the algorithm collects a minute video for Instagram [5]. If the focus is to spectacularly announce the extension of a contract—in a couple of clicks one can receive the best moments of the player's career from WSC Sport.

Naturally, it was possible to do the same thing before, but manually and with a constant risk of missing something. Israel's technology, which is rapidly learning itself and already conquering other sports, is a perfect example of intensive development when a standardized task is transferred to smart development, and human resources can be directed to creativity [5]. In addition, manual assembly takes extra time, and WSC Sport returns the advantage to clubs and the league in the fight for the second screen of the viewer.

There is no exact data or research details on how this algorithm works, but the scientists say their technology is constantly learning, so the further they go, the better they can harness the full power of artificial intelligence [5]. The algorithm prioritizes in real time because different events can be the most important in different matches. The most important thing is not just to collect moments in a cut, but with the help of AI technologies to succinctly formulate the same plot that has developed in the match [2].

WSC Sport is one of the flagships in the creation of ultra-personalized content. In addition to the NBA, the American football league MLS, the most progressive sports media in the world, [2], and others are already cooperating with the company. They all understand that in parallel with global coverage, the demand for individual media consumption is growing. Among the audience, many want to determine for themselves what exactly and when exactly to watch [6]. Instead of targeting content to specific markets or demographics, one can now give users the chance to customize it for themselves [2]. Integration with a chatbot in a messenger is unlimited access for a regular user to a huge content database, which is automatically (or rather, automatically) created for any occasion. Now any media company and organization that wants to be active and visible through their content has a powerful additional tool.

Yet, the WSC Sports technology seems to be too complex and is based on a substantial data-gathering process. A new solution has been suggested by [1]. Scholars have utilized the contrastive learning methodology. The main focus of their framework is to teach the algorithm an unsupervised scope of actions that can be applied for multiple downstream tasks. The AI is trained to create a random mapping, localization, identification, and segmentation of images in a video stream. The algorithm also performs the cropping and editing functions presenting the best images in a given sequence thereby transforming the original frame to avoid spoilers and increase attention and awareness of the specific product. The scholars have initially considered their methodology with a focus on contrastive learning to conduct the visual transformation in the most straightfor-

ward way possible. By applying the dropout sequence as part of the contrastive learning procedure, [1] have managed to develop an unsupervised sequence embedded in the algorithm that is now capable of a general transformation with the highlight presentation as the outcome. Scholars have developed an unsupervised framework that is capable of picking the most attractive clips that further make up a single vector. Then the dropout sequence is applied for the transformation to take place and to be able to map the two vectors, localize the moment of interest, and cluster the imaging, with its further segmentation. As presented in Figure 1, the framework allows us to cluster the video based on content, yet with the dropout fluctuation and change. This allows for better clusters to be chosen for further highlighting and production of video embedding. Figure 1 shows the same video content embedded in vectors with the non-highlights (left) cluster that have been dropped, and the highlights (right) cluster that the algorithm has decided to retain for further operation and editing.

Figure 1 clearly shows that the non-highlight clusters do not fit the framework as they overlap, are inconsistent, and are different in both scope and content. Each video feed is presented as a different color, with the application of random dropout sequences to categorize the feed. This demonstrates the difficulty there is in directly utilizing the video feed and using it in the video highlights compilation from the start.

3. Suggested Approach and Results

3.1. Overview

The approach presented in [1] suggests first splitting the video feed into same-length clips. Each clip is represented using a vector, whereas the whole video is presented in the form of a clip sequence. Then the C3D action recognition model is implemented to extract the raw features of dimension. The main goal here is to achieve high scores through a contrasting learning framework (Figure 2).

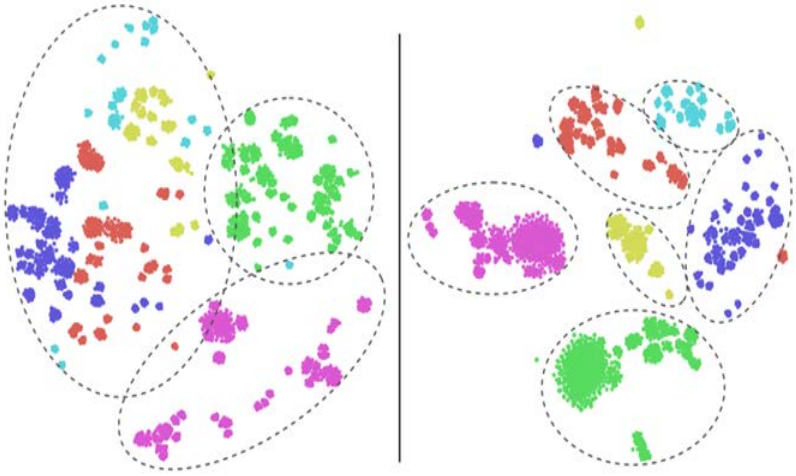

Figure 1. The non-highlights clusters vs highlights clusters [1].

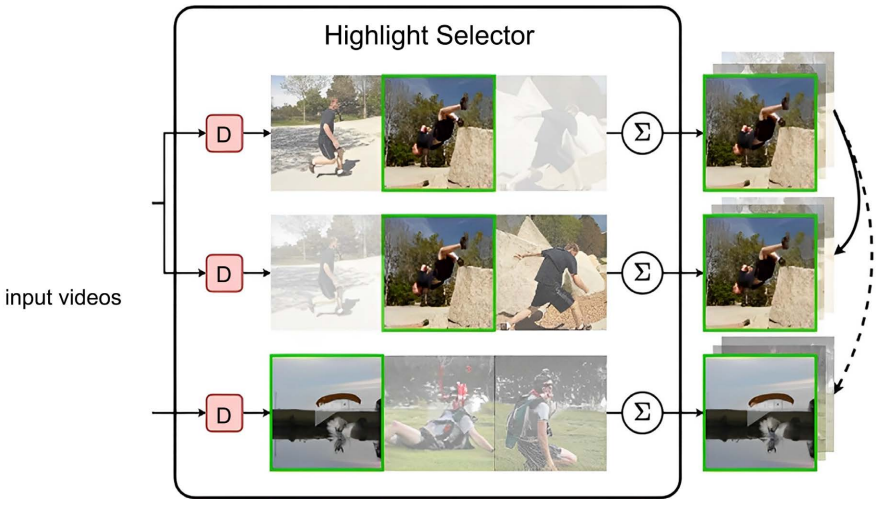

Figure 2. C3D action recognition clip selector.

Here we learn to map the changed vectors farther away from each other in the sequence. The clip selector establishes a specific attention score further used to compute a specific sum of the clips measured against the scale of the highlight clips.

3.2. Generic Attention Layer + Highlight Selector

The generic attention layer is an important factorial inclusion to make the notation as simple as possible. **Figure 3** demonstrates the peculiar impact that this layer has on the transformation process of the entire perception of the video feed. The attention layer is defined as such that queries Y using X. At the output, we receive a generic attention score in the form of an attention map with a particular interdependence level between x_i and y_i. This allows us to determine the interaction and the level of interchange between the various sequences presented in the video feed. This is also required for the smooth transition of the audio feed between the transformed vectors. After all, if the query and key highlight sequences are different in length, that would mean that there would be a residual part requiring us to remove the clip from the framework altogether (as in **Figure 1**).

The Farcana video feed from the Twitch clip was broken into an equal number of clips with a predetermined number of frames for each feed. As a result, we obtained N clips. As has been mentioned each clip is passed through a pre-trained C3D action recognition model, to size each clip. The generic attention layer then models the relationship between clips to determine their coherency and their compatibility. Then the second sequence of generic attention layer identification is conducted to determine the specific features of the query pool. As a result, we receive a set of attention scores with a comparison to the highlights score. These scores remain unsupervised in the process of embedding with the application of contrasting learning. The highlight detection is aimed at capturing the relationship between different clips, even if they are not from the same moment in time of the Farcana Twitch feed.

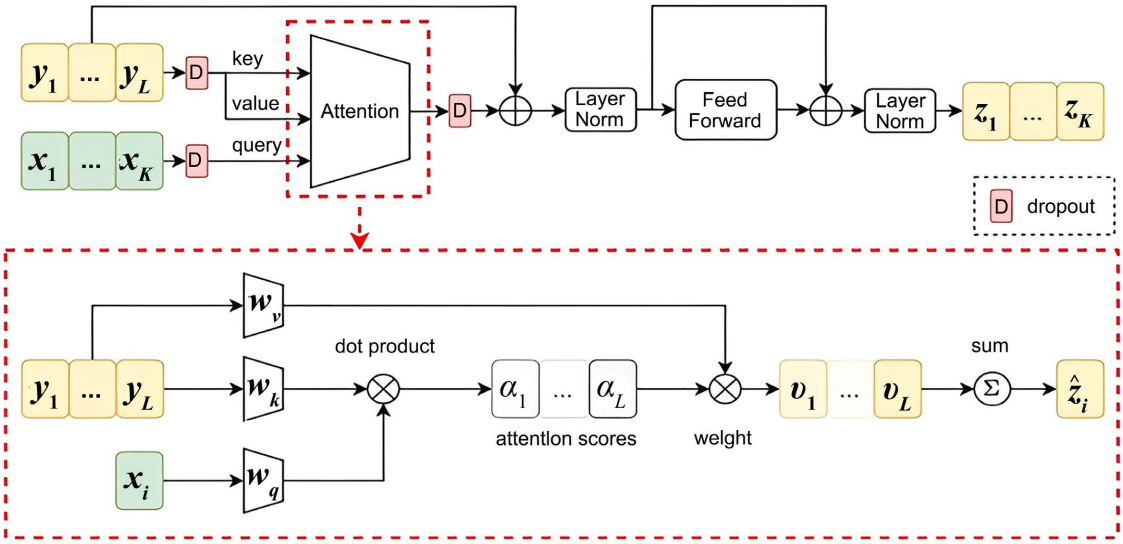

Figure 3. Generic attention layer.

3.3. Analysis

The approach in [1] has allowed to teach the algorithm the process of video clustering bearing in mind the ongoing dropout fluctuations. This model learns to pick out moments that cannot be considered highlights due to the peculiar dropout moment and a low generic attention score. Yet, if the model does foremost drop the non-highlight clips, it may have been safe to assume that the remaining would be safe to use for the highlights. Yet, this is not the case, as should the algorithm stop at this moment, the quality of clusters will deteriorate, and therefore the performance of the contrastive learning model suggested by [1] would simply become irrelevant and inapplicable in any potential sphere of operation.

Instead, one suggests resetting the algorithm from start by comparing the non-highlight clips to the remaining hypothetically highlight clips. We use the C3D Action recognition model on both, having both groups of clips pass the whole process again to receive the output of the first generic action level. Again, the remaining clips are sampled into non-highlights and highlights. This cycle is repeated another twenty times with a different set of dropout requirements. The embedding outputs should form a separate cluster, with a specific intra-cluster distance. The closer the video embeddings are to each other the greater interrelationship between them. Whereas a greater distance would demonstrate that the videos are too far apart from each other both in terms of content and even generic attention scores. To measure the distance, the cosine distance is used. This score allows the algorithm to receive intra- and inter-cluster statistical data and consequentially compute the mean value across all videos.

The highlight clips form a tight cluster with embeddings located close to each other, whereas those at a greater intra-cluster distance identify the clip to be a non-highlight clip. In addition, as is visible from **Figure 1**, the highlight clips form separated clusters that are located at a substantial difference from each

other. This allows the algorithm to pinpoint the best moments and further edit them into a highlight clip.

4. Conclusion and Limitations

This paper has utilized the model presented in the study by [1] to determine the possibility of attaining highlights from the Farcana 2.0 version Twitch video feed. We have implemented a simple unsupervised method that is repeatable and does not require additional complex state-of-the-art equipment or method. This model distinguishes the dropouts from other models, with a highlight received as the outcome. No external large data is collected, with the model having the possibility to remain unsupervised/weakly supervised in highlight detection.

Despite the evident success, there are certain limitations to the process. It has been determined that the non-highlight clips do not contain the same amount of information, yet there is still a chance that this can be disproven over a higher number of sequences running the model. A possible explanation could be the substantially different content of the videos. A possible solution is by choosing video feeds that are closest to the specific video in question to form a topical constitution of negative pairs. The current study only chose random feeds of Farcana 2.0 version release gameplay.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Badamdorj, T., Rochan, M., Wang, Y. and Cheng, L. (2022) Contrastive Learning for Unsupervised Video Highlight Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, 18-24 June 2022, 14042-14052. https://doi.org/10.1109/CVPR52688.2022.01365

[2] Hong, F.T., Huang, X., Li, W.H. and Zheng, W.S. (2020) Mini-Net: Multiple Instance Ranking Network for Video Highlight Detection. *European Conference on Computer Vision*, Springer, Cham, 345-360. https://doi.org/10.1007/978-3-030-58601-0_21

[3] Jiao, Y., Li, Z., Huang, S., Yang, X., Liu, B. and Zhang, T. (2018) Three-Dimensional Attention-Based Deep Ranking Model for Video Highlight Detection. *IEEE Transactions on Multimedia*, **20**, 2693-2705. https://doi.org/10.1109/TMM.2018.2815998

[4] Campbell, C., Sands, S., Ferraro, C., Tsao, H.Y.J. and Mavrommatis, A. (2020) From Data to Action: How Marketers Can Leverage AI. *Business Horizons*, **63**, 227-243. https://doi.org/10.1016/j.bushor.2019.12.002

[5] Liu, W., Yan, C.C., Liu, J. and Ma, H. (2017) Deep Learning-Based Basketball Video Analysis for Intelligent Arena Application. *Multimedia Tools and Applications*, **76**, 24983-25001. https://doi.org/10.1007/s11042-017-5002-5

[6] Ye, Q., Shen, X., Gao, Y., Wang, Z., Bi, Q., Li, P. and Yang, G. (2021) Temporal Cue-Guided Video Highlight Detection with Low-Rank Audio-Visual Fusion. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 7950-7959. https://doi.org/10.1109/ICCV48922.2021.00785