25th Computer Vision Winter Workshop
Domen Tabernik, Alan Lukežič, Klemen Grm (eds.)
Rogaška Slatina, Slovenia, February 3–5, 2020

movie2trailer: Unsupervised trailer generation using Anomaly detection

Orest Rehusevych

The Machine Learning Lab of
Ukrainian Catholic University, ELEKS Ltd.
Lviv, Ukraine
rehusevych@ucu.edu.ua

Taras Firman

Ukrainian Catholic University, ELEKS Ltd.
Lviv, Ukraine
firman@ucu.edu.ua

Abstract. *In this work, we present movie2trailer - a novel unsupervised approach for automatic movie trailer generation. To our knowledge, it is the first-ever application of anomaly detection to such a creative and challenging part of the trailer creation process as a shot selection. One of the main advantages of our approach over the competitors is that it does not require any prior knowledge and extracts all needed information directly from the input movie. By leveraging the recent advancements in video and audio analysis, we produce high-quality movie trailers in equal or less time than professional movie editors. The proposed approach reaches state-of-the-art in terms of visual attractiveness and closeness to the "real" trailer. Moreover, it exposes new horizons for researching anomaly detection applications in the movie industry. The trailers, that were used in evaluation stage are available at the following link - https://bit.ly/2GbOj4R.*

1. Introduction

With the massive expansion of online video-sharing websites such as YouTube, Vimeo, and others, movie promotion through advertisements becomes much more widespread than earlier. In contrast to previous decades, nowadays, trailers became the most crucial part of the movie promotion campaign. Since the trailer creation requires a lot of human efforts and creative decisions considering the selection of scenes, montages, special effects, teams of professional movie editors have to go through the entire film multiple times to select each potential candidate for the best moment. This process can take between 10 days to 2 years to complete [27]. On the high-cost movies, there can be up to six different trailer creation companies involved in this process. During working on the creation of a movie trailer, the editor makes multiple alternative versions of the trailer, the best to be chosen by the target group of specialists afterward. According to [27], there can be created up to 200 variants of the trailer for the target movie. These facts reveal what a significant role a trailer plays in movie success and how much resources it takes to produce a great trailer.

All these factors were the main stimulus for us to make a research on the problem of automatic trailer generation and raise its possibilities to an entirely new level. In our opinion, the area of automatic trailer generation has not been explored enough, and many people underestimate the capabilities of AI advancements over the recent years and how they could be utilized to create high-quality trailers similar to the real one. We strongly believe that AI, to some point, can simulate the expertise and creativity of professional movie editors and reduce huge costs and time consumption.

2. Related works

In this section, we present a short overview of all main approaches for movie trailer generation. The literature divides these methods into two main groups: fully-automated methods and those with human assistance. Until the advent of advanced methods, video summarization techniques, such as Clustering-based Video Summarization [9] and Attention-based Video Summarization [19], were applied to the problem of automatic movie trailer generation. Because of this fact, all the approaches, which focus on movie trailer generation, were using video summarization techniques as competitors in the evaluation stage. Similarly to them, as an addition, we compare our approach with Muvee[1] - commercial video summarization software.

2.1. Video2Trailer (V2T)

Vid2Trailer (V2T) [12] is a content-based movie trailer generation method. In this paper, the authors set two main requirements for trailers properties to be pleased: they must include specific symbols, such as the title logo sequence shot or/and the main theme music, and they should be visually and audibly attractive to the viewers. As is stated, the algorithm satisfies both of them. The complete pipeline consists of three main stages: symbol extraction, impressive components extraction, and reconstruction. According to the authors, at the time of the publication in 2010, V2T was more appropriate to trailer generation than conventional movie summarization techniques.

2.2. Point Process-Based Visual Attractiveness Model (PPBVAM)

In [29], the authors propose an automatic trailer generation approach, which mainly focused visual attractiveness. Based on common observation, authors assume that during attractive scenes, viewers mostly look at the same area of the screen and, on the other side, lost their focus when boring scenes appear. Consequently, they propose a surrogate measure of visual attractiveness based on viewers' eye-movement, named fixation variance, which is further used as a metric for shots selection. To sum it up, in this paper, authors propose the novel metric for visual attractiveness named fixation variance and learn an attractiveness dynamics model for movie trailers by applying self-correcting point process methodology [13, 22]. The authors mention that their approach outperforms all the previous automatic trailer generation methods and reaches SOTA in terms of both efficiency and quality.

2.3. Human-AI joint trailer generation

Unlike the two automatic trailer generation algorithms mentioned above, IBM Research, in cooperation with 20th Century Fox, introduced the system for first-ever Human-AI trailer creation collaboration, described in [26]. The primary purpose of the system was to identify ten candidates among all movie scenes as the best moments. Further, the professional filmmaker would edit and arrange these moments to construct a comprehensive movie trailer. The system was designed to understand and encode patterns of emotions presented in horror movies. The following steps were performed: Audio Visual Segmentation, Audio Sentiment Analysis, Visual Sentiment Analysis, Scene Composition Analysis, Multimodal Scene Selection. The main system advantage is that it can significantly reduce the involvement of the film editor in the trailer creation process.

3. Approach

Based on our assumptions that by using anomaly detection we can reveal the nonstandard frames among others and that they are the ones that are regularly used in professional movie trailers, we have created a system for automatic trailer generation without any previous knowledge about the target movie. One of the main advantages of our approach is its flexibility in terms of visual appearance. By changing visual features, we can easily put accents on what a user wants to observe in the generated trailer. Figure 1 shows the high-level architecture of our approach.

3.1. Shot Boundary Detection

Shot boundary (transition) detection is one of the major research areas in video signal processing. The main problem it solves is the automated detection of changes between shots in the video. Even though cut detection appears to be an easy task for a human, it is still a non-trivial task for machines. Taking into account a vast number of different types of transitions during shot changes, the problem remains very challenging even nowadays. A lot of researches [14, 30, 1] studying a comparison of various shot boundary detection algorithms were made. Still, there is no silver bullet for detecting all types of transitions accurately. For our work, we decided to go with an open-source Python library for detecting scene changes in videos and automatically splitting the video into separate clips, named *PySceneDetect* [3]. It provides us with two different detection methods:

- Simple threshold-based fade in/out detection
- Advanced content-aware fast-cut detection

The second one appeared to be more appropriate for our problem. The content-aware scene detector finds areas where the difference between two subsequent

[1] https://www.muvee.com

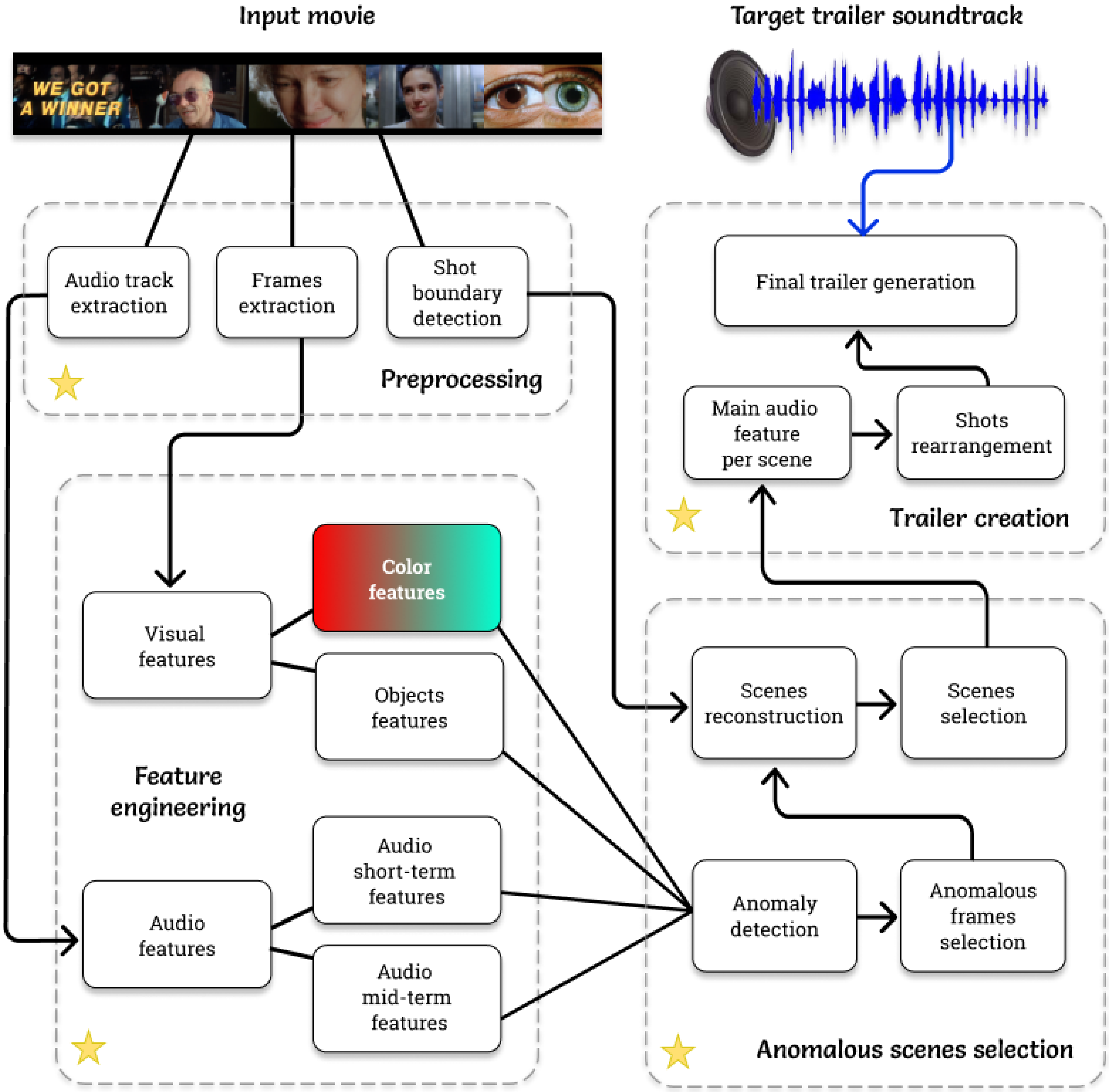

Figure 1: High-level architecture of movie2trailer.

frames exceeds the set threshold value. In contrast to the most traditional scene detection methods, the content-aware detector allows detecting cuts between the scenes, both containing similar content. With a fine-tuned threshold, this approach can detect even minor and sudden changes, such as jump cuts.

3.2. Feature engineering

Feature engineering without exaggeration can be named the most important part of the whole pipeline. This component directly influences the outcomes of all further steps and consequently changes the appearance of the final generated trailer. The selection choice of features leads to changes in what exactly a person wants to see in a trailer. For example, if we want to have a lot of scenes with explosions in our trailer, we need to add a custom feature, which is responsible for detecting explosions (can be done either with the video or audio feature). Table 1 shows all three types of features (visual, audio short-term, and audio mid-term) that was calculated for the given movie.

3.2.1 Visual features

Visual features were selected based on our understanding of what people usually expect to see in the trailer. They can be divided into two subgroups: color model features and object detection features. For color features, we chose the HSL color model,

Visual	**Audio short-term**	**Audio mid-term**
Delta hue	Zero Crossing Rate	Mean and standard deviation of all 34 audio short-term features
Delta saturation	Energy	
Delta lightness	Entropy of Energy	
Content value	Spectral Centroid	
Number of people	Spectral Spread	
Number of non-people objects	Spectral Entropy	
Total number of objects	Spectral Flux	
Area of detected people	Spectral Rolloff	
Area of detected non-people	MFCCs	
Total area of detected objects	Chroma Vector	
	Chroma Deviation	

Table 1: The chosen visual, audio short-term and audio mid-term features.

where H corresponds to hue, S - saturation, L - lightness. These properties represent a color spectrum in different forms, which we consider an essential visual aspect of human perception. Additionally, we include the *content value* parameter (mean between Hue, Saturation, and Lightness) to this group of features, as it takes the most significant role in our shot boundary detection process. Hence we are inclined to believe that content value provides information responsible for shot change detection. All the other visual features can be attributed to another (object detection) group. The creation of these features was achieved by leveraging the capabilities of Faster R-CNN [23], pretrained on MS COCO dataset [16]. As a result, we were able to distinguish 80 classes of the most common objects, such as a person, different vehicles, various animals, and everyday things in their natural context. From the extracted information about objects on the frame, we construct six features which can be split into quantity and area groups. The first one was taken because of the hypothesis that frames with many people correspond to scenes with lots of action which keeps viewers' attention on the screen. Another group was formed under the assumption that close-up shots are attractive to view.

3.2.2 Audio short-term and mid-term features

In the majority of the cases, the most salient audio parts are accompanied by outstanding visual scenes and vice versa. Therefore, audio features are not less important than the visual ones. In our algorithm we have used a set of audio features previously introduced in [6]. All audio features were retrieved by exploiting the potential of the open-source library for audio signal analysis named *pyAudioAnalysis* [6]. The main reason of this choice was that because of the significant coverage of sound signal properties, these features had been used in multiple audio analysis and processing techniques. Before the feature extraction step, an audio signal is usually cut into nonoverlapping windows (frames). For the short-term feature sequences, we have used a frame size of 50 msecs of an audio signal and a 1-second window size for the mid-term, correspondingly. As a result of feature extraction, we get a sequence of 34-dimensional and 68-dimensional feature vectors for short-term and mid-term audio signals, respectively. Mid-term features accumulate statistics over the short-term features for a more extended time period to catch more general changes in the audio signal. The statistics include the mean and variance over each short-term feature sequence. To sum it up, we have gathered together all the essential properties of the audio signal for both time and frequency domains that could be further utilized for multiple purposes: from detecting speech among other sounds, to determining the saliency of different parts of the audio.

3.3. Anomalous scenes selection

Anomalous scenes selection is a long process containing multiple steps: anomaly detectors selection, retrieval of anomaly frames for each type of features, choice of abnormal visual frames, audio short-term and mid-term frames, merging them together taking into an account the difference in duration of each feature type frame, constructing final set of video frames, scenes reconstruction, threshold-based anomalous scenes selection. Figure 2 shows the complete pipeline of this scene selection approach.

3.3.1 Anomaly detectors selection

Having extracted frame-level visual, short-term, and mid-term audio features for the entire movie, now we are ready to use them in the process of anomalous scenes selection. As a first step, we need to determine what anomaly detectors to use. Based on our experiments, we have concluded, that by applying multiple types of detectors, the result would be much more credible than by using a single one, because of the very different underlying logic between all of them, the various types of data that generated features were based on and possibly very different scale of features. For that reason, we have chosen 8 anomaly detection algorithms covering most of these cases, which could be divided into 4 groups (2 detectors per each group):

- **Linear models: MCD** (Minimum Covariance Determinant) [24], **OCSVM** (One-Class SVM) [20].
- **Proximity-based models: LOF** (Local Outlier Factor) [2], **HBOS** (Histogram-based Outlier Score) [7].
- **Ensembles: IsoForest** (Isolation Forest) [17], **Feature Bagging** [11].
- **Neural networks: AE** (fully-connected AutoEncoder) [10], **MO-GAAL** (Multiple-Objective Generative Adversarial Active Learning) [18].

3.3.2 Anomalous frames selection

With the selected anomaly detectors, we run them separately on each type of the features: visual, audio short-term, and audio mid-term. Each of these types includes its own set of features with diverse frame duration. Since each of the detectors has its pros and cons, we have introduced a voting system to determine the most appropriate frames of each feature type. The frame is considered suitable if at least five of eight detectors have chosen it as anomalous. Having selected frames of each feature type, we reduce audio short-term and mid-term frames to their corresponding visual frames taking into consideration the duration periods of each feature group frame. After that, We obtain a set of anomalous final video frames by taking an intersection between all groups of frames. These frames serves as the basis to identify trailer-worthy scenes from an input movie.

3.3.3 Scenes selection and reconstruction

With the already defined final set of visual frames and information about each scene start and end timestamps, we are ready to reconstruct scenes. The primary constraints for scenes selection are that scenes should have the maximum percentage of anomalous frames and their total duration should be not less than the length of the accompanying soundtrack. Through the visual examination of selected scenes, we have determined that the scenes with the highest number of abnormal frames are the most valid candidates for making the trailer.

3.4. Shots rearrangement

Shots reordering is a beneficial step because it can additionally improve the overall human perception of the viewed video by maximizing the attractiveness with some particular order of shots. By conducting multiple experiments, we have tested a hypothesis that lots of percussive timbres (claps, snares, drums) accompany fast shots with lots of action. Furthermore, we had an assumption that there are some audio features, that should be responsible for detecting percussive sounds. Based on the idea, described in [8], we have found out that by using zero-crossing rate, we could be able to detect such type of sounds quickly and accurately. With our experience watching numerous trailers, we have concluded that in most trailers, the accompanied music increases its intensity through the entire video. To validate that idea, we have calculated the zero-crossing rate vector for each scene and tried different flows with sorting by mean, median, max value of this feature. After that, we have visually examined each of the generated trailers and compared them with the trailer, where scenes are ordered as in the original movie. Since the visual appearance of the arranged by audio feature trailers was visually worse than the one ordered by chronology, we consequently stuck to the latter option.

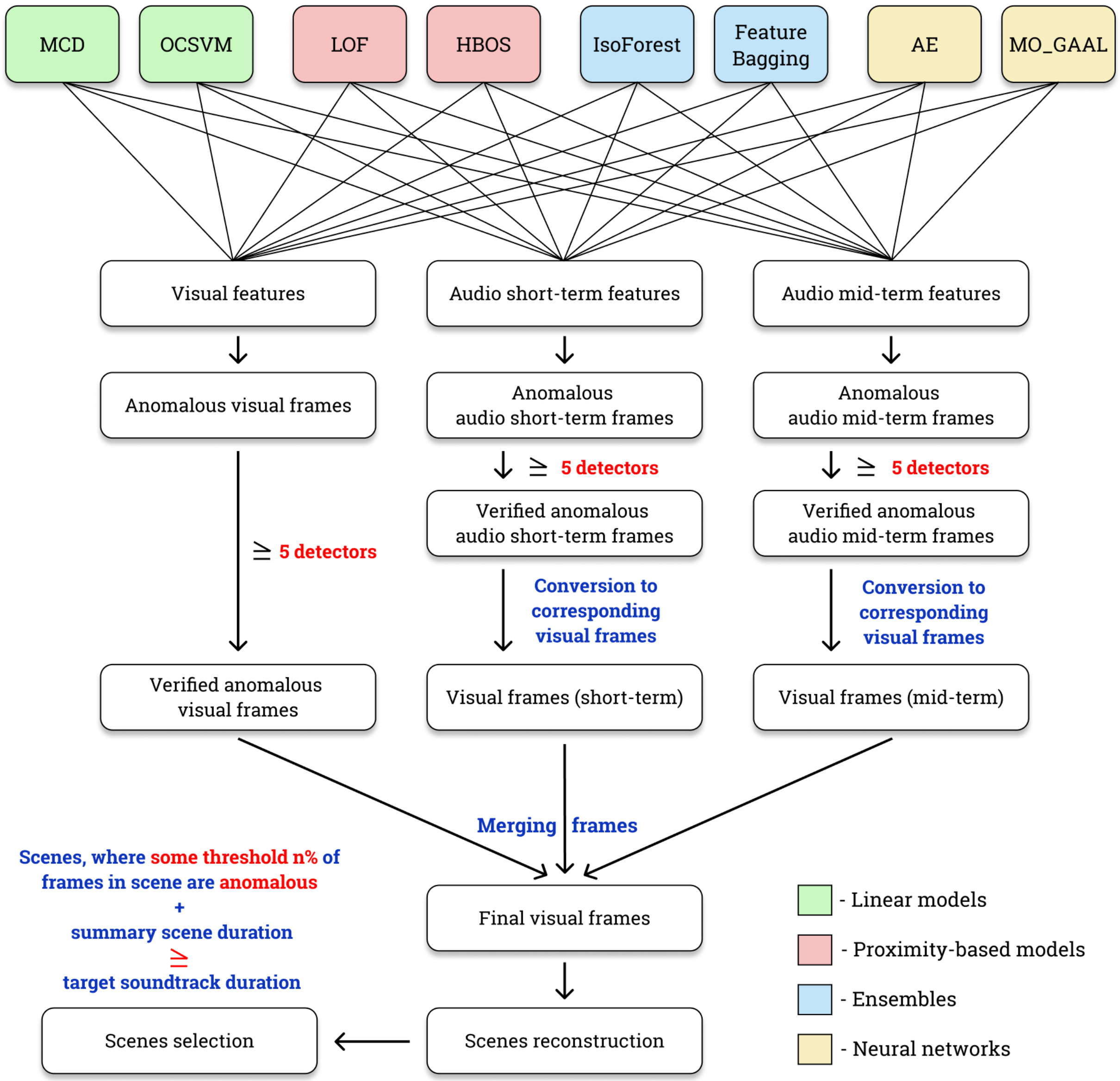
MCD
OCSVM
LOF
HBOS
IsoForest
Feature Bagging
AE
MO_GAAL
Visual features
Audio short-term features
Audio mid-term features
Anomalous visual frames
Anomalous audio short-term frames
Anomalous audio mid-term frames
≥ 5 detectors
≥ 5 detectors
≥ 5 detectors
Verified anomalous audio short-term frames
Verified anomalous audio mid-term frames
Conversion to corresponding visual frames
Conversion to corresponding visual frames
Verified anomalous visual frames
Visual frames (short-term)
Visual frames (mid-term)
Merging frames
Scenes, where some threshold n% of frames in scene are anomalous
+
summary scene duration
≥
target soundtrack duration
Final visual frames
Scenes selection
Scenes reconstruction
🟩 - Linear models
🟥 - Proximity-based models
🟦 - Ensembles
🟨 - Neural networks

Figure 2: The detailed pipeline of anomalous scenes selection.

4. Evaluation and results

In this section, we evaluate our method **movie2trailer** against all the leading opponents for the automatic trailer generation problem:

- **V2T** [12] - Trailer generation method;
- **Muvee** - Commercial software for video summarization;
- **PPBVAM** (Point Process-Based Visual Attractiveness Model) [29] - SOTA for automatic trailer generation;
- **RT** - The original official real trailers;
- **RTwS** - The same real trailers without speech information

4.1. Qualitative results

For the objective evaluation, we have taken a series of measures to avoid assessment bias. None of the volunteers has seen any of the generated trailers previously. None of the volunteers knew the order of the methods while observing trailers. All final generated trailers were downscaled to the resolution of other trailers (480x360) produced by our competitors' approaches, and all the speech pieces were replaced with the original soundtrack. Similarly to our predecessors, on the input, we give the entire movie without cutting any parts from it to remove spoilers. With the steps above, we can be confident that all the approaches are on an equal footing and would be evaluated without any bias. Similarly to [12] and [29], we have invited 23 volunteers with different movie tastes and preferences to evaluate the visual appearance of each testing trailer created with different approaches by answering on the following three questions:

- **Appropriateness:** "How similar this trailer looks to an actual trailer?"
- **Attractiveness:** "How attractive is this trailer?"
- **Interest:** "How likely you are going to watch the original movie after watching this trailer?"

For each question, a volunteer should give an integer score of how much he/she agree on the particular statement on the Likert scale [15]: from 1 (the lowest) to the 7 (the highest). Figure 3 shows the overall results for all 3 testing movies: *"The Wolverine (2013)", "The Hobbit: The Desolation of Smaug (2013)", "300: Rise of an Empire (2014)"*. Authors of the **PPBVAM** provided trailers [2] generated with main competitors' approaches for abovementioned movies. We were limited to use only these three movies since the reproduction of some parts of competitors' algorithms is infeasible. The results of the poll show that our method is superior to **V2T**, **Muvee** and **PPBVAM** in all three questions, indicating that our approach to shot selection using anomaly detection is reasonable, and can provide us with such types of shots that satisfy our subjective feelings and perception.

We believe that **RTwS** and **RT** were usually preferred more by volunteers, because all trailers generated using automatic trailer generation methods were deprived of speeches, subtitles, and special effects of montages. Since the information that these factors provide to improving visual attractiveness, we should also supplement our system with these information sources in the future.

4.2. Quantitative results

To the best of our knowledge, the only publicly available method for video aesthetics assessment - Semi-automatic Video Assessment System [21]. This framework incorporates diverse set of visual features that are closely related to aesthetics: Luminance, Optical Flow, Colourfulness and lots of other. All these features are used by SVM [4] to determine the level of aesthetics and interestingness of

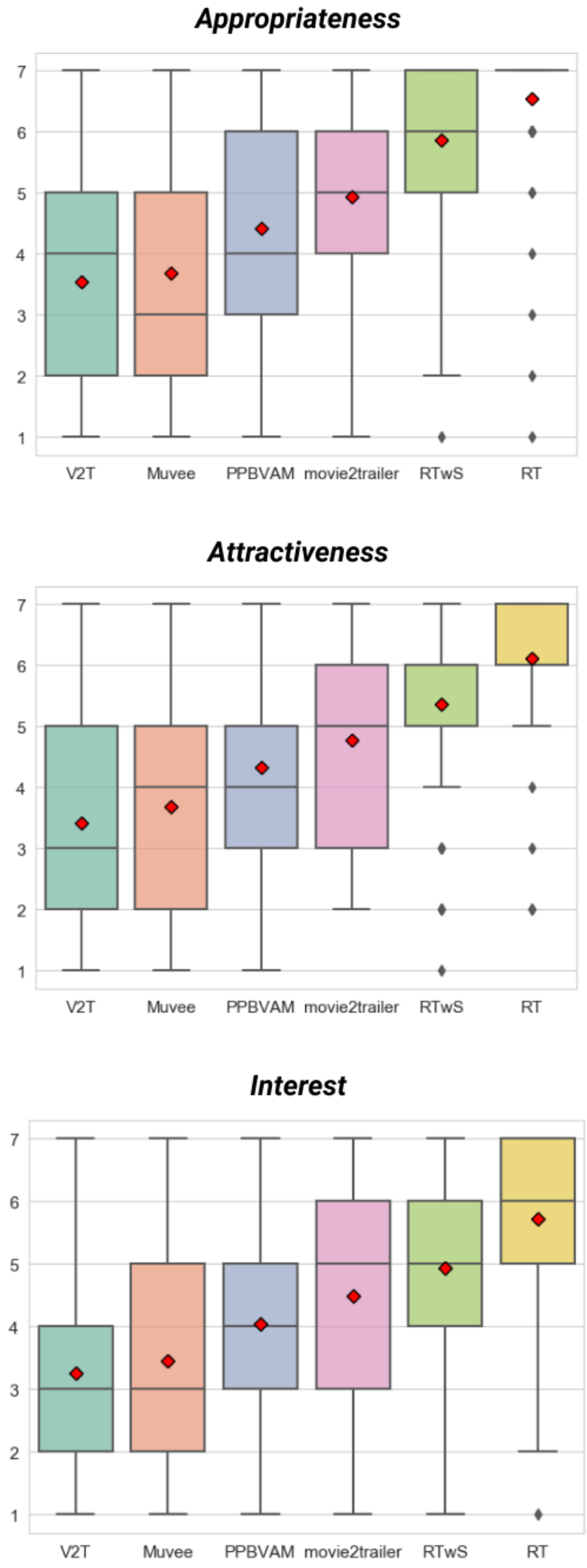

Figure 3: The box plots of scores for various methods on three questions considering Appropriateness, Attractiveness and Interest. The dark lines inside boxes are medians and red diamonds are means. Dark points outside of the whiskers are outliers.

[2] https://vimeo.com/user25206850/videos

metric	V2T	Muvee	PPBVAM	**ours**	RTwS	RT
mean	4.39	4.31	4.00	**4.73**	4.71	**4.76**
std	0.42	0.33	0.36	0.60	0.65	0.62

Table 2: Quantitative statistics of the NIMA scores.

the target video. To train that method, CERTH-ITI-VAQ700 dataset [28] was used. In view of its large size, the authors decided to use only 1 second of each video, and as a result, their method does not work well on longer videos. During our evaluation, it gave aesthetics score 0 for all trailers, including real and generated ones.

We propose a new approach for video aesthetics evaluation based on evaluating the aesthetics of each video frame separately:

1. Extract all the frames f_i from the video.
2. Compute aesthetics score s_i for each frame f_i.
3. Compute metrics (mean, standard deviation) based on the obtained aesthetics scores s_i.

As a candidate for image aesthetics scoring function, we tested NIMA (Neural Image Assessment) [5] and Will People Like Your Image? [25].

The results obtained using NIMA aesthetics scores (from 1 to 10) (Figure 4 and Table 2) shows that our approach works at the level of **RT** (real trailer).

We have also applied the same approach for evaluation using another image aesthetics assessment algorithm [25]. We have not included the results of this scoring method in this section, because in all cases, it evaluated real trailers worse than the generated ones.

The quantitative results obtained by the aesthetics scoring systems shows that our method outperforms all existing automatic movie trailer generation approaches and is at the level of the real trailers.

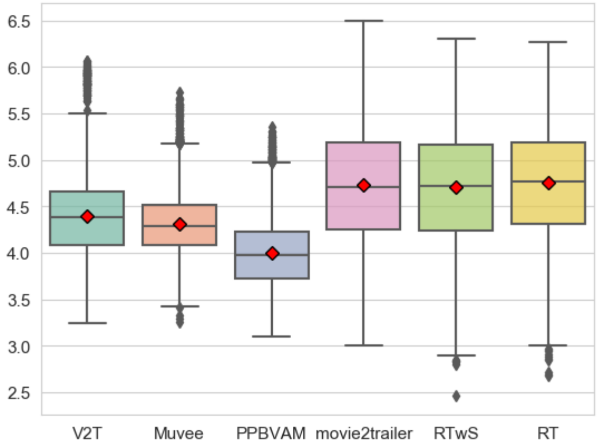

Figure 4: Quantitative comparison of movie trailer approaches based on NIMA aesthetics metric [5].

5. Conclusion

In this paper, we have presented an unsupervised trailer generation method, named *movie2trailer*. Our approach automatically creates high-quality trailers by identifying anomalous frames relying on the selected set of visual and audio features. A series of quantitative and qualitative experiments show that *movie2trailer* outperforms all the previous automatic trailer generation methods in terms of visual attractiveness and similarity to the "real" trailer and thus is more appropriate to trailer generation than previous techniques. We demonstrated the tremendous potential of the intelligent multidomain analysis system in applying to such a profoundly creative task as creating a movie trailer. This research study opens doors for further investigations of the anomaly detection applications in the movie industry.

Acknowledgements

The authors were supported by the ELEKS Ltd. and Ukrainian Catholic University. Special thanks to Oles Dobosevych for his help with the preparation of this publication and providing computational resourses, without which this project hasn't been possible.

References

[1] S. H. Abdulhussain, A. R. Ramli, M. I. Saripan, B. M. Mahmmod, S. A. R. Al-Haddad, and W. A. Jassim. Methods and challenges in shot boundary detection: A review. *Entropy*, 20(4):214, 2018. 2

[2] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA.*, pages 93–104, 2000. 5

[3] B. Castellano. Video scene cut detection and analysis tool. https://github.com/Breakthrough/PySceneDetect, 2018. 2

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 7

[5] H. T. Esfandarani and P. Milanfar. NIMA: neural image assessment. *IEEE Trans. Image Processing*, 27(8):3998–4011, 2018. 8

[6] T. Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12), 2015. 4

[7] M. Goldstein and A. Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59–63, 2012. 5

[8] F. Gouyon, F. Pachet, O. Delerue, et al. On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy*, 2000. 5

[9] A. G. Hauptmann, M. G. Christel, W. Lin, B. Maher, J. Yang, R. V. Baron, and G. Xiang. Clever clustering vs. simple speed-up for summarizing rushes. In *Proceedings of the 1st ACM Workshop on Video Summarization, TVS 2007, Augsburg, Bavaria, Germany, September 28, 2007*, pages 20–24, 2007. 1

[10] G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 3–10, 1993. 5

[11] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844, 1998. 5

[12] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Automatic trailer generation. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 839–842, 2010. 2, 6, 7

[13] V. Isham and M. Westcott. A self-correcting point process. *Stochastic Processes and their Applications*, 8:335–347, 1979. 2

[14] R. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases VII, San Jose, CA, USA, January 26-29, 1999*, pages 290–301, 1999. 2

[15] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 7

[16] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755, 2014. 4

[17] F. T. Liu, K. M. Ting, and Z. Zhou. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 413–422, 2008. 5

[18] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, and X. He. Generative adversarial active learning for unsupervised outlier detection. *CoRR*, abs/1809.10816, 2018. 5

[19] Y. Ma, L. Lu, H. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of the 10th ACM International Conference on Multimedia 2002, Juan les Pins, France, December 1-6, 2002.*, pages 533–542, 2002. 1

[20] L. M. Manevitz and M. Yousef. One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154, 2001. 5

[21] P. Martins and N. Correia. Semi-automatic video assessment system. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, CBMI 2017, Florence, Italy, June 19-21, 2017*, pages 33:1–33:7, 2017. 7

[22] Y. Ogata and D. Vere-Jones. Inference for earthquake models: A self-correcting model. *Stochastic Processes and their Applications*, 17:337–347, 1984. 2

[23] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. 4

[24] P. J. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999. 5

[25] K. Schwarz, P. Wieschollek, and H. P. A. Lensch. Will people like your image? learning the aesthetic space. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 2048–2057, 2018. 8

[26] J. R. Smith, D. Joshi, B. Huet, W. H. Hsu, and J. Cota. Harnessing A.I. for augmenting creativity: Application to movie trailer creation. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1799–1808, 2017. 2

[27] C. Snyder. How movie trailers are made - business insider. https://www.businessinsider.com/how-movie-trailers-are-made-2018-7, 2018. 1

[28] C. Tzelepis, E. Mavridaki, V. Mezaris, and I. Patras. Video aesthetic quality assessment using kernel support vector machine with isotropic gaussian sample uncertainty (KSVM-IGSU). In *2016 IEEE International Conference on Image Processing, ICIP 2016,*

Phoenix, AZ, USA, September 25-28, 2016, pages 2410–2414, 2016. 8

[29] H. Xu, Y. Zhen, and H. Zha. Trailer generation via a point process-based visual attractiveness model. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 2198–2204, 2015. 2, 6, 7

[30] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang. A formal study of shot boundary detection. *IEEE Trans. Circuits Syst. Video Techn.*, 17(2):168–186, 2007. 2