UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

movie2trailer: Unsupervised trailer generation using Anomaly detection

Author:
Orest Rehusevych

Supervisor:
PhD Taras Firman

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2019

Declaration of Authorship

I, Orest Rehusevych , declare that this thesis titled, “movie2trailer:
Unsupervised trailer generation using Anomaly detection” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Everything I learned, I learned from the movies."

Audrey Hepburn

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

movie2trailer:
Unsupervised trailer generation using Anomaly detection

by Orest Rehusevych

Abstract

In this work, we present a novel unsupervised approach for automatic trailer generation - movie2trailer. 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 can 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. An example of generated with our approach trailer for the movie *"Requiem for a dream"* can be observed on *YouTube*.

Acknowledgements

First of all, I want to express my immense gratitude to my supervisor Taras Firman, for his guidance throughout all the research, for generating a key idea of our approach, lots of recommendations considering every component of the project and always giving valuable feedback. Also, want to thank Oles Dobosevych for providing computational resources (GeForce GTX 1080) without which this project hasn't been possible. Special thanks to Anton Tarasov and Oleh Smolkin for exciting discussions considering possible improvements for the proposed approach. I am very thankful to Ukrainian Catholic University for the knowledge I have gained through all these 4 years. Also, I'm sincerely grateful to Yarynka Lutsyk for believing in me more, than myself. Finally, I want to thank my parents for their permanent reminders concerning the writing of the thesis, which kept me in shape and not allowed to forget about it for a single day.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
DL	**D**eep **L**earning
ANN	**A**rtificial **N**eural **N**etwork
SOTA	**S**tate-**O**f-**T**he-**A**rt
SVM	**S**upport **V**ector **M**achine
PCA	**P**rincipal **C**omponent **A**nalysis
GAN	**G**enerative **A**dversarial **N**etwork
SBD	**S**hot **B**oundary **D**etection
MS COCO	**M**icro**s**oft **C**ommon **O**bjects in **Co**ntext
MFCCs	**M**el **F**requency **C**epstral **C**oefficients
MCD	**M**inimum **C**ovariance **D**eterminant
OCSVM	**O**ne-**C**lass **SVM**
LOF	**L**ocal **O**utlier **F**actor
HBOS	**H**istogram-**b**ased **O**utlier **S**core
AE	**A**uto**E**ncoder
FCN	**F**ully **C**onvolutional **N**etwork

List of Symbols

std	standard deviation
cov	covariance matrix
$dct, idct$	DCT transformation pair

Dedicated to my supportive parents

Chapter 1

Introduction

1.1 Motivation

With the massive expansion of online video-sharing websites such as YouTube, Vimeo, and others, movie promotion through advertisements becomes much more accessible than earlier. In contrast to previous decades, nowadays, trailers became the most crucial part of the movie promotion campaign. These days, people get no less excited for a movie trailer as they do for the actual movie. The quality and visual attractiveness of the trailer directly influence the box office success of the movie. Because of that fact, film studios are ready to spend an enormous amount of money and human resources to create such trailers that would maximize the number of viewers at the movie theaters. 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 anywhere between 10 days to 2 years to complete. On the high-cost movies, there can be up to six different trailer creation companies involved in this process. All these factors were the main reasons for us to make a research on the problem of automatic trailer generation and push its possibilities to a completely new level. In our opinion, the area of automatic trailer generation hasn't been explored enough, and a lot of 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 the professional movie editors and reduce huge costs and time consumptions.

1.2 Major requirements for a great movie trailer

In order to understand the key requirements of the movie trailers, first, we need to find out what makes a movie trailer great. First of all, the trailer takes the role of the basic premise of the movie. It gives an overall idea of what we should expect from the movie, but not reveal any of the plot twists or provide hints that could lead to them. Successful trailers find the golden mean between giving too much and not enough information. Secondly, the good trailers introduce the audience with only some main characters. Instead of bombarding viewers with the whole cast of the movie, the successful trailers focus only on a few central characters. Thirdly, as long as the accompanying music reflects the spirit of the film, it is a way to majesty. The last and the most important one - catch the mood of the movie. Keep attention to emotions rather than plot points. Make the trailer speak from itself, and the audience would be excited. These are the primary requirements for a great movie trailer.

1.3 The process of trailer creation

Since trailer creation requires so many human efforts, let's deconstruct this process on some standard components. First of all, goes the creation of the concept and the structure of the future trailer. Then movie editor should go through the entire footage to determine the best candidates for each of the previously selected storylines. With having already chosen scenes, editors start the montage. This part is the most extensive since it includes the selection of transitions between shots, applying visual effects, putting voice-overs in the most appropriate places, adding music and sound effects. The abovementioned are just the most basic steps of trailer creation. Numerous components (such as color editing, synchronizing of shot changes with accompanying percussive sounds, etc.) can be applied additionally to improve the overall look of the video. 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. These different variants of trailers can count up to 200 per movie. This fact reveals what a significant role a trailer plays in movie success and how much resources it takes to produce a great trailer.

Chapter 2

Related works

2.1 Video summarization

Trailer generation problem has a long history that takes a start from video summarization techniques. The main difference between them is that video summarization aims to pick the minimum number of frames or shots in order to preserve the storyline and the goal of trailer generation is to produce high-satisfactory videos with the maximum level of attractiveness while not revealing main plot twists. Since these two problems lie in the same domain and are highly interconnected between each other, video summarization methods were widely applied to movie trailer generation before more advanced techniques. The standard pipeline (Figure 2.1) for video summarization consists of video segmentation, feature extraction, redundancy detection based on extracted features, and finally, generation of video summary.

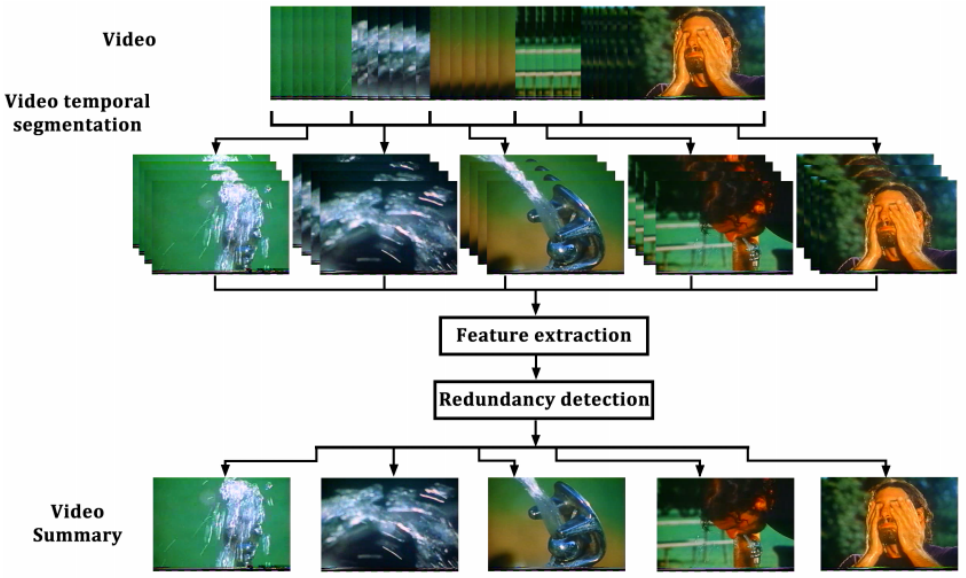

FIGURE 2.1: A general approach for video summarization.
Source: [Cahuina and Chávez, 2013]

The most obvious examples of video summarization techniques that were applied to movie trailer generation are Attention-based Video Summarization [Ma et al., 2002] and Clustering-based Video Summarization [Hauptmann et al., 2007]. Recent advancements in DL, precisely the key idea behind GANs [Goodfellow et al., 2014], allowed applying Generative-Adversarial framework to Unsupervised Video

Summarization [Mahasseni, Lam, and Todorovic, 2017]. Moreover, the very different types of networks now are applied to Video Summarization, such as Attention-Based Encoder-Decoder networks with LSTMs [Ji et al., 2017], Foveated convolutional neural networks [Wu et al., 2018], Semantic Attended Networks [Wei et al., 2018], and others. Considering the fact that videos produced by video summarization techniques are frequently not attractive enough to become a movie trailer, follows the conclusion that the problem of movie trailer generation should be solved with different than video summarization approaches.

2.2 Movie trailer generation

2.2.1 Existing fully-automated approaches

Vid2Trailer: Automatic Trailer Generation

Vid2Trailer (V2T) [Irie et al., 2010] 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 (Figure 2.2) consist of three main stages: *symbol extraction, impressive components extraction, and reconstruction.*

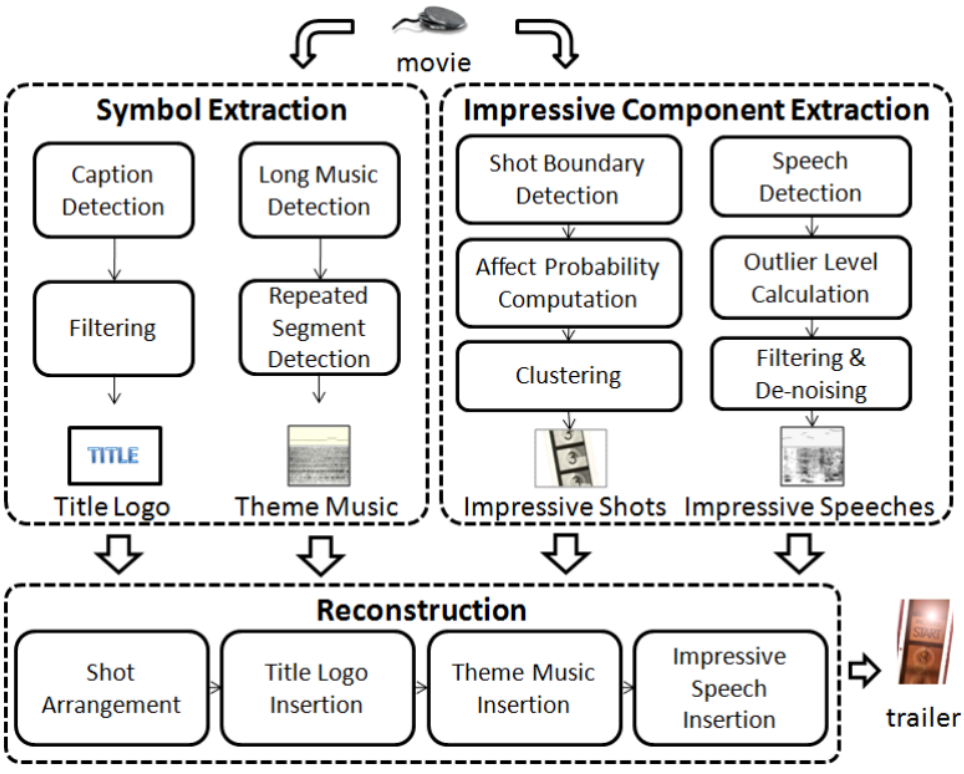

FIGURE 2.2: Overview of Vid2Trailer (V2T). Source: [Irie et al., 2010]

Symbol extraction

Symbol extraction step consists of the title logo and theme music extraction. The first one is performed due to the observation that most often title logo differs from other text captions (e.g., the cast, production staff, and the distributors) by largest font-size. Based on this idea, the caption detection algorithm proposed in [Kuwano et al., 2000] is applied to the first 10% of the movie and title logo is extracted from detected captions. Given the assumptions that the theme music lasting more time than the other music components, and the melody of theme music is used repeatedly throughout the movie; the most often-used music segment with the most extended duration suppose to be the theme music. Retrieving the longest music segments from the original audio track is made by applying the music detection method presented in [Minami et al., 1998]. Then, the most often-used melody detection process based on melody matching is performed by extracting a 12-dimensional *chroma vector* [Goto, 2006] from all the extracted music segments. Further, pair-wise similarities between all vectors and repeated scores are calculated, and finally, the music segment, which achieves the highest repeated score, has been selected as the theme music.

Impressive components extraction

Impressive components (both speech and video segments) extraction from the original visual and audio sources is done using an *affective content analysis* technique. In order to analyze the affects of movie segments, the approach, described in [10] is used. This method estimates eight-dimensional probability vector p(ei | xi), where ei and xi are the affect and the shot-level feature of the i-th shot respectively, and each dimension corresponds to the probability of Plutchik's eight basic emotions [Plutchik, 2001]. The calculation of affective probability vector is done by combining latent Dirichlet allocation (LDA) [Blei, Ng, and Jordan, 2001] and conditional probability table (CPT). Finally, representative affective shots are retrieved by applying the clustering, specifically, *affinity propagation (AP)* [Frey and Dueck, 2007]. Initially, all speeches are extracted from the source audio by speech extraction [Minami et al., 1998]. Because "impressive" speeches are used to be deeply emotional, their audio features are expected to be very different from the "average" ones. Therefore such speeches are obtained by using outlier detection, specifically by applying *Gaussian mixture* [Zhuang et al., 1996] and estimating its parameters with the *Expectation-Maximization (EM)* algorithm [Dempster, Laird, and Rubin, 1977]. As a final post-processing step, spectral-subtraction-based de-noising is applied to all of the selected speeches.

Reconstruction

Reconstruction consists of shot arrangement and component insertion parts. The proposed method estimates the impact of the shot sequence and formulates shots re-ordering as an optimization problem which aims to maximize the impact on the viewers. The approach for estimating the *affective impact* of a shot sequence is formed due to the idea of *Bayesian Surprise* [Itti and Baldi, 2005] - a framework for calculating the surprise level induced in the viewers through the observation of visual information. As the concluding step, V2T inserts title logo sequence shot, main theme music, and impressive speeches into an already rearranged shot sequence. Since for these elements, there are some apparent locations where they should be put, the component insertion process is rule-based and consist of three simple rules (one per each component), that can be observed in Figure 2.3.

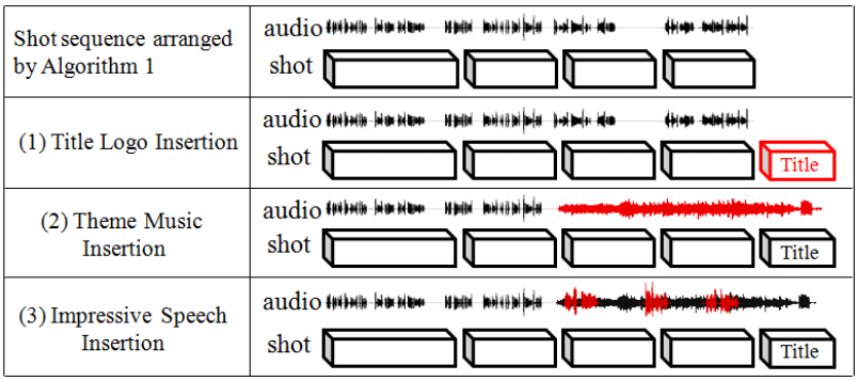

FIGURE 2.3: Flow of component insertion. Source: [Irie et al., 2010]

According to the authors, at the time of the publication in 2010, V2T was more appropriate to trailer generation than conventional techniques for trailer generation such as Clustering-based Video Summarization [Hauptmann et al., 2007] and Attention-based Video Summarization [Ma et al., 2002].

Trailer Generation via a Point Process-Based Visual Attractiveness Model

In [Xu, Zhen, and Zha, 2015], the authors propose an automatic trailer generation approach, which is densely connected with visual attractiveness. The system consists of three key elements:

- An empirically validated measure of trailer attractiveness;
- An efficient algorithm for selecting and re-arranging shots in order to maximize the level of attractiveness;
- A useful method for synchronizing shots and music for upgraded viewer experience.

Measure of attractiveness

Based on common observation, authors make an assumption 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*. The following experiments empirically validated the effectiveness of this metric. Fourteen volunteers were invited to watch eight movie trailers (containing 1083 shots). During this process, the motion of their gazes was traced using Tobii T60 eye tracker, and further, the mapped fixation points in each frame were calculated. The authors define the fixation variance as the determinant of the covariance matrix of the fixation points:

$$\sigma_j^{(i)} = \det(\mathrm{cov}([\mathbf{x}_j^{(i)}, \mathbf{y}_j^{(i)}])),$$

where i - denotes the ordinal number of the shot, j - the ordinal number of the frame, $[\mathbf{x}_j^{(i)}, \mathbf{y}_j^{(i)}]$, where $\mathbf{x}_j^{(i)}, \mathbf{y}_j^{(i)} \in^{14}$ - the vector of the coordinates of the fixation points of the 14 volunteers. Furthermore, the fixation variance was averaged for the frames belonging to the same shot. In order to validate the hypothesis that the averaged fixation variance reflects the spread of attention while watching the shot, two types of shots were manually labeled, and the statistics of their fixation variances were

calculated. The results in Table 2.1 shows that the mean and median of the fixation variance of boring shots has large values and vice versa. Hence, the fixation variance is negatively correlated with visual attractiveness. Moreover, it was found that even though the attractiveness within one shot decreases over time, it increases with the montages between shots.

	mean(σ)	median(σ)	variance(σ)
Boring shots	1.19	0.45	0.03
Attractive shots	**0.60**	**0.22**	0.01

TABLE 2.1: The statistics of normalized fixation variance ($\times 10^8$). Source: [Xu, Zhen, and Zha, 2015]

Similarly to visual attractiveness, the authors have found that the dynamics of music is also highly correlated with the montages between shots. The algorithm described in [Hou, Harel, and Koch, 2012] was used to retrieve the saliency curve of a music piece:

$$\hat{\mathbf{m}} = \mathrm{G}((\mathrm{idct}(\mathrm{sign}(\mathrm{dct}(\mathrm{m}))))^2)$$

where dct($\cdot$) and idct($\cdot$) are a DCT transformation pair, sign($\cdot$) is the function that returns 1, -1, and 0 for positive, negative, and zero inputs, respectively and G($\cdot$) is a Gaussian filter. The peaks determined after resampling $\mathbf{m}$ with the number of frames were considered as the saliency points of the music. The correct indicator of a music-montage match was valid if a montage appeared $\pm$ 6 frames within a peak timestamp. The experiments on the 8 sample trailers showed that the saliency points are highly correlated with those of the montages. Figure 2.4 presents an example that verifies the aforementioned:

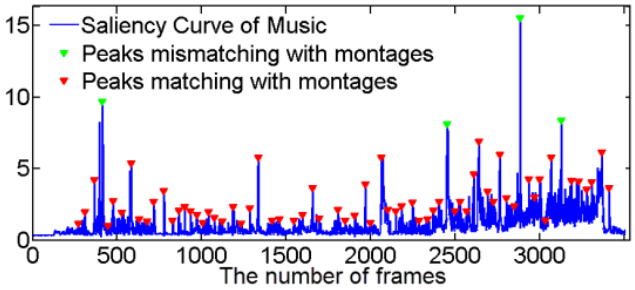

FIGURE 2.4: The Saliency Points of Music vs Montage Positions (Trailer of "The Bling Ring"). Source: [Xu, Zhen, and Zha, 2015]

Point Process-based Attractiveness Model

Their modeling assumption is that the fixation variance is highly correlated with the number of viewers losing their attention on the screen, which directly connects the notion of the attractiveness of a video with a specific point process model. Although they do not observe the event sequence directly, assuming that the fixation variance is proportional to the number of viewers losing their attention. As a result, they defined the self-correcting point process, that is a point process with some intensity function. The intensity function of the self-correcting point process increases exponentially with some rate. To summarize, they have defined the attractiveness model as a self-correcting point process with a global intensity function for a time interval.

Trailer Generation

The complete process of trailer generation includes three main steps: *candidate selection*, *parameter assignment call to enumerate environment*, *graph-based stitching*.

The first step is provided just for selecting shots that satisfy pre-defined constraints. In general, this step is required as hard filtering of not useful shots. Next step helps to select some additional parameters that might affect the final result of model selection. The final decision they build is based on solving a complicated combinatorial optimization problem. The overall idea is based on finding the solution of recursive influence on the selection of subsequent shots. Finally, they reduced this problem to solving shortest path [Dijkstra, 1959] in some graph G.

To sum it up, in this paper, authors propose the novel metric for visual attractiveness and learn an attractiveness dynamics model for movie trailers by applying self-correcting point process methodology [Isham and Westcott, 1979], [Ogata and Vere-Jones, 1984]. The authors mention that their approach outperforms all the previous automatic trailer generation methods and reaches SOTA in terms of both efficiency and quality.

2.2.2 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 [Smith et al., 2017]. The primary purpose of the system was to identify 10 candidates among all movie scenes as the best moments. Further, the professional filmmaker would edit and arrange these moments in order to construct a comprehensive movie trailer. The high-level architecture of the complete process can be observed in Figure 2.5.

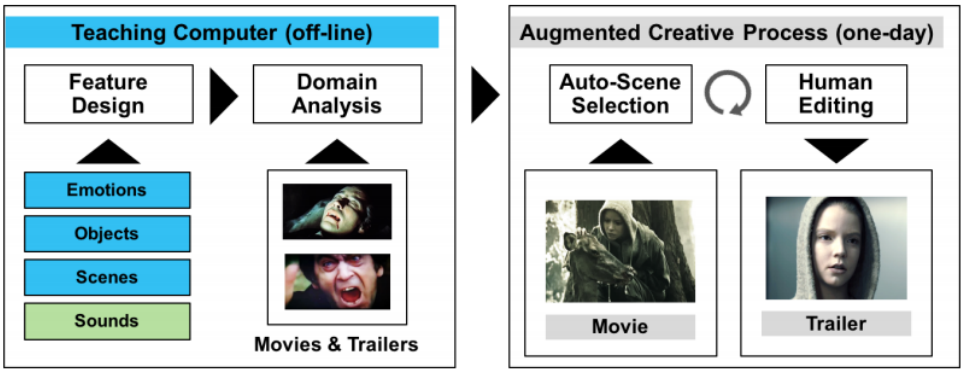

FIGURE 2.5: The high-level architecture of the Intelligent Multimedia Analysis driven Trailer Creation Engine.
Source: [Smith et al., 2017]

The system was designed to understand and encode patterns of emotions presented in horror movies. In order to do that, the following steps were performed: *Audio Visual Segmentation*, *Audio Sentiment Analysis*, *Visual Sentiment Analysis*, *Scene Composition Analysis*, *Multimodal Scene Selection*.

Audio Visual Segmentation

The first step in reaching an extensive understanding of a movie was to divide it into audio/visual snippets. After shot boundary detection and splitting process, each shot was represented with a visual key-frame for further feature extraction. Audio segmentation was done using OpenSmile framework [Eyben et al., 2013]. As both audio and visual segments were retrieved individually, in order to form a full visual sentiment feature, visual features for all key-frames that lies within an audio segment were aggregated together.

Audio Sentiment Analysis

An audio analysis was performed using OpenSmile [Eyben, Wöllmer, and Schuller, 2010], and particularly, audio emotion recognition was done using OpenEAR extension [Eyben, Wöllmer, and Schuller, 2009]. A probability scores for each emotion that corresponds to a model class predictions were assigned with respect to some input audio signal. Except discrete features mentioned above, two continuous dimensional features, named valence and arousal were retrieved. Together all these features form an 18-dimensional emotion vector representing a complete audio sentiment feature.

Visual Sentiment Analysis

There was a need to create a complete representation over emotions spectrum in order to gain a full understanding of the movie scene visual sentiment structure. The authors used Sentibank [Borth et al., 2013] to retrieve visual sentiment information from movie key-frames. At first, an image is classified into an adjective-noun pair category. Then the sentiment scores specific to that category are assigned. A visual sentiment feature is constructed using the 24 sentiment scores (based on 24 emotions from the Plutchik's wheel of emotions [Plutchik, 2001]) corresponding to the sentiment distribution for the highest ranked adjective-noun pair. To create a composite visual sentiment representation for a whole segment, the authors calculated a dimension-wise mean across all the key-frames in the segment.

Scene Composition Analysis

The authors observed that the atmosphere and other aesthetic factors are often affected by specific scene composition rules. Places 205 CNN model [Zhou et al., 2014] was chosen for modeling scene visual attributes because the network actually categorizes the types of locations and emphasizes the global context in scene composition.

Multimodal Scene Selection

With already extracted high-level visual and audio features, the created AI system analyzed all the retrieved data and computed a statistical model, that captures the outstanding characteristics from trailers from the specific movie genre (in this case - horror). As a part of feature analysis, Principal Component Analysis (PCA) was applied to the extracted features. The three main dimensions were considered to capture the characteristics peculiar to the horror genre. This assumption was taken as a basis to distinguish good movie scenes of the same genre. As a result, each movie scene was projected onto 3-dimensional space, then the scenes with the highest response were chosen to be the best candidates for trailer creation. During an audio-visual inspection of selected scenes, authors noticed that each of the three major dimensions corresponded in some way to scare, tender, and suspense.

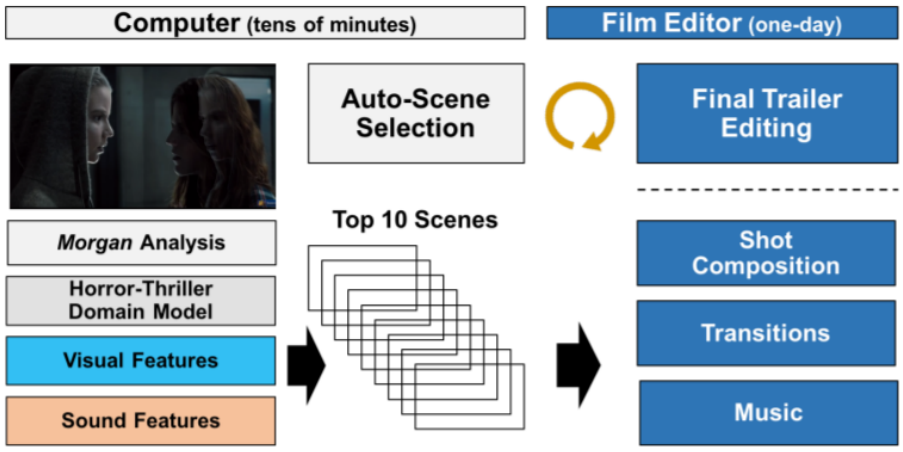

FIGURE 2.6: The roles played by the computer and the human in the augmented creative trailer making process.
Source: [Smith et al., 2017]

To sum it up, the roles of the Augmented Intelligence system and the film editor can be inspected in Figure 2.6. So, the primary system responsibilities are a comprehensive analysis of the target movie, exploration of the genre domain, visual and audio features extraction, features interpretation, automatic scene selection. The main system advantage is that it can significantly reduce the involvement of the film editor in the trailer creation process. Moreover, this system is an excellent example of how AI can be applied to such a highly creative task, as movie trailer generation.

Chapter 3

Background information

3.1 Anomaly detection algorithms

Anomaly detection refers to the task of identification observations, items or events that do not conform to the standard, expected behavior by differing significantly from the majority of the data [Chandola, Banerjee, and Kumar, 2009]. These observations have different names: anomalies, outliers, exceptions, surprises, novelties considering the domain where the anomaly detection is applied. Anomaly detection nowadays is actively applied in multiple areas [Ahmed, Mahmood, and Hu, 2016], such as Cyber-Intrusion detection, System health monitoring, Surveillance, Sensor networks, Image processing, and a lot of other domains.

Types of anomalies

A critical aspect of choosing anomaly detection technique is the nature of the given anomaly. A detailed analysis of the taxonomy [Huang, 2018] of the types of anomalies gives us a possibility to categorize anomalies into four specific classes (Figure 3.1):

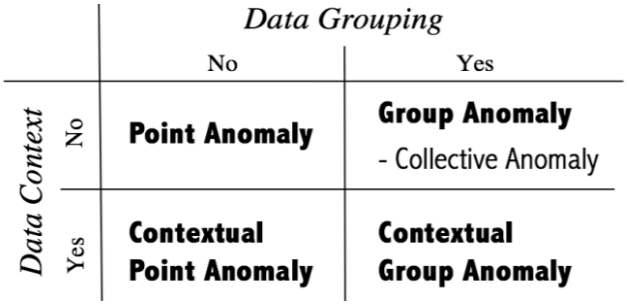

FIGURE 3.1: The taxonomy of anomalies. Source: [Huang, 2018]

- ***Point anomaly*** is a single observation that deviates notably from all the observations according to some predefined criteria.
- ***Group anomaly*** is a collection of observations that are grouped based on a predefined criterion, that does not fit into the standard patterns of other sets of observations.
- ***Contextual point anomaly*** is a single observation that deviates notably from all the observations according to some predefined criteria under a specific context.
- ***Contextual group anomaly*** is a set of observations that are grouped based on a predefined criterion, that does not follow the usual patterns of other sets of observations under a specific context.

Types of anomaly detection algorithms based on input

Taking into account the existence of the labels to our input data, anomaly detection approaches can be divided into the following three categories:

Supervised anomaly detection. In this setup (Figure 3.2), applied techniques assume the availability of labeled training and testing data with normal and abnormal classes. Typically, such approaches are similar to traditional pattern recognition, where we have a classification problem that differentiates regular data from the unusual ones. The main difference is that in the case of anomaly detection, we tackle the problem of imbalanced class distributions. The other additional issue is obtaining accurate, and representative labels, especially for the abnormal class, is still very challenging. Taking into account the two problems mentioned above, not all classification approaches are suitable for this task. Support Vector Machines (SVM) or Artificial Neural Networks (ANN) should be a good option for solving anomaly detection in the supervised mode.

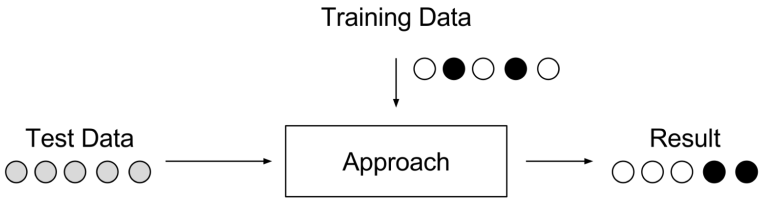

FIGURE 3.2: Supervised anomaly detection Source: [Kibish, 2018]

Semi-Supervised anomaly detection. In this mode (Figure 3.3), applied techniques assume the availability of labeled training data for only normal data or only abnormal ones as the input. Since such techniques do not require labels for both classes, they are more applicable than supervised ones. The typical approach used in these techniques is to create a model for only the class corresponding to the normal behavior, and then use it to find anomalies in the test data. Since a model contains information about a single concept, semi-supervised anomaly detection can be represented as a one-class classification problem. Due to the single type of data samples, one-class classification gets rid of the problem of imbalanced class distributions. The main problem that remains is that the input data instances cannot be inaccurate or noisy in order to one-class classification methods be robust. The most common approaches used in a semi-supervised mode are One-class SVMs and Autoencoders.

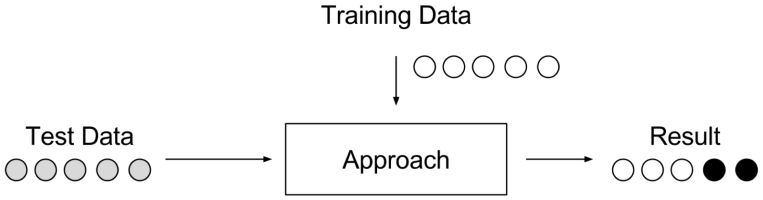

FIGURE 3.3: Semi-Supervised anomaly detection.
Source: [Kibish, 2018]

Unsupervised anomaly detection. Approaches that operate in this mode (Figure 3.4) does not require any prior knowledge of data (no label data is presented), thus are most widely applicable. These techniques have to analyze the data in order to infer or to make an assumption of the concept of abnormality. Usually, such techniques make an implicit assumption that normal instances are considerably more frequent than anomalies in the data. If this hypothesis is a wrong one, these techniques greatly suffer from high false alarm rate. Typically, in an unsupervised mode, densities or distances are used to evaluate what is normal and what is an outlier.

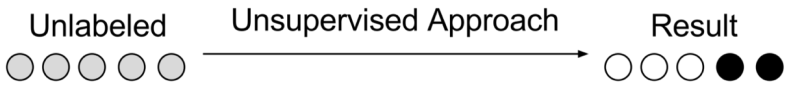

FIGURE 3.4: Unsupervised anomaly detection. Source: [Kibish, 2018]

Most of the popular and conventional unsupervised anomaly detection approaches [Goldstein M, 2016] are shown in Figure 3.5.

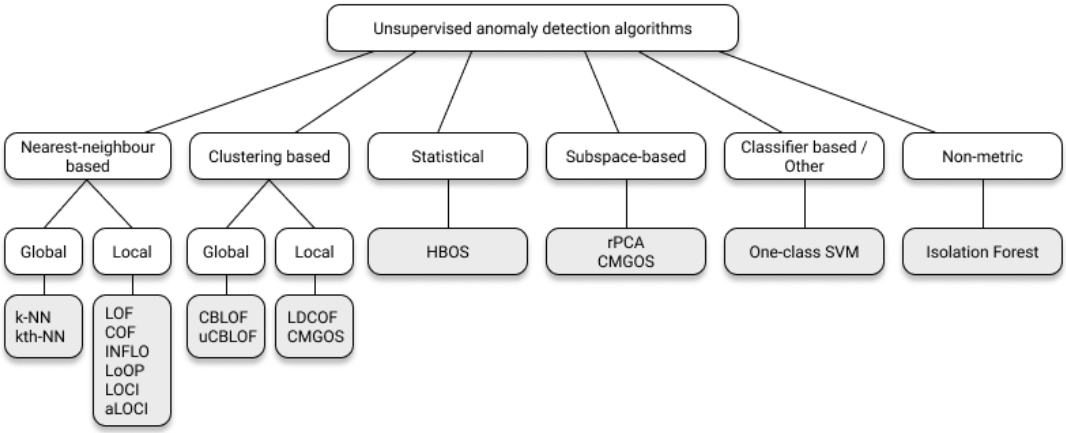

FIGURE 3.5: Unsupervised anomaly detection approaches classification diagram.

Types of anomaly detection algorithms based on output

Based on the output of the anomaly detection algorithm, they can be grouped into two typical types of solutions:

- ***Continuous scores*** are more preferred to be used when there is a demand for a detailed analysis of the data.
- ***Discrete labels*** greatly simplify the anomaly detection system design and are more convenient for end users to interpret results.

Chapter 4

Proposed approach

Based on our assumptions that by using anomaly detection we can reveal the non-standard 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. The high-level architecture of our approach can be inspected in Figure 4.1.

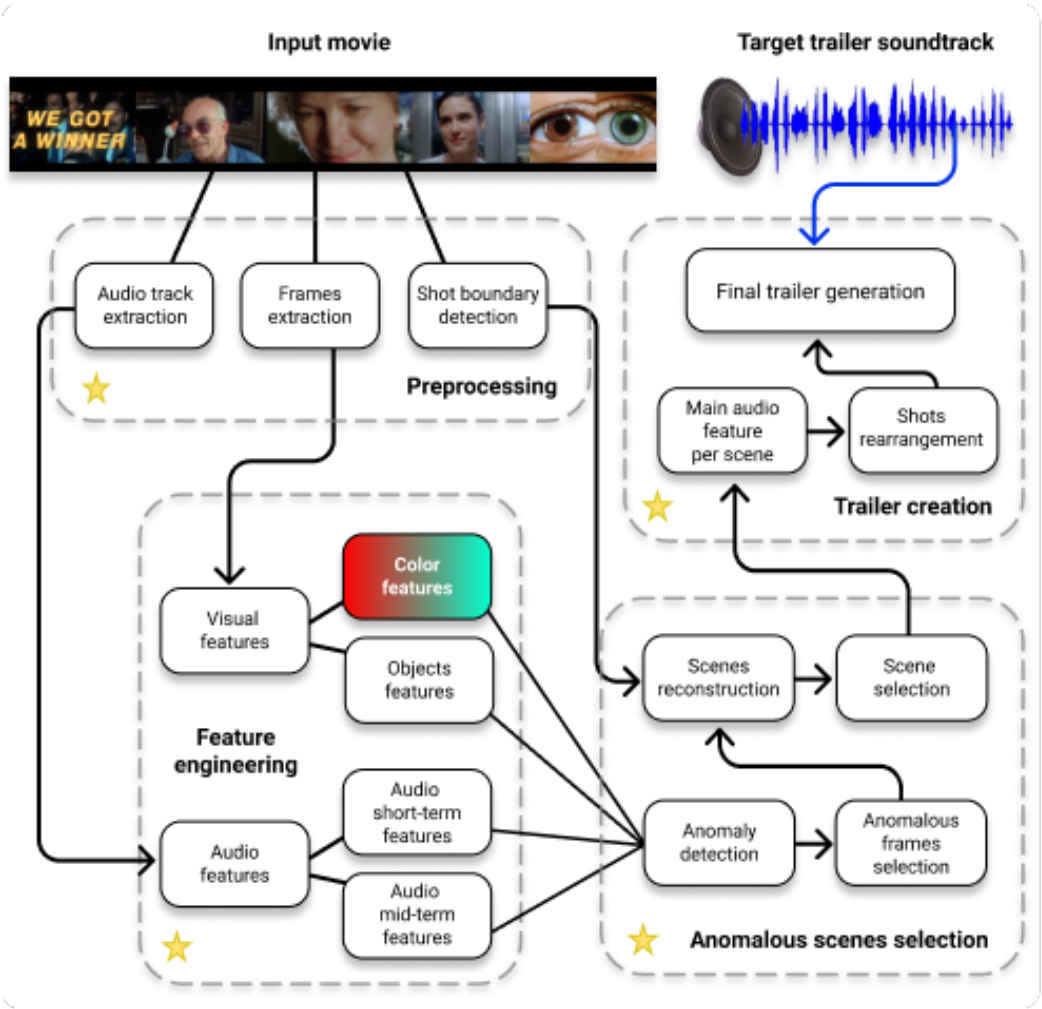

FIGURE 4.1: High-level architecture of movie2trailer.

4.1 Shot boundary detection

Shot boundary (transition) detection is one of the major research areas in video signal processing. The main problem its 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. If each frame of the video were supplemented with some additional information, such as the position of the camera and time when each frame was taken, that would become a non-trivial task to determine video shot changes for a computer. 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 [Lienhart, 1999], [Yuan et al., 2007], [Abdulhussain et al., 2018] studying a comparison of various shot boundary detection algorithms were made, and still, there is no silver bullet for detecting all types of transitions accurately. For our work, we decided to go with open-source Python library for detecting scene changes in videos and automatically splitting the video into separate clips, named *PySceneDetect* [Castellano, 2018].
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 frames exceeds the set threshold value. Rather than most of the traditional scene detection methods, the content-aware detector allows detecting cuts between the scenes both containing some content. With fine-tuned threshold, this approach can detect even such minor and sudden changes, such as jump cuts (Figure 4.2). By applying this method to our target movie *"Requiem for a dream"*, we retrieved 2061 shots.

FIGURE 4.2: An example of jump cut on 2 adjacent shots.
Source: [*Breathless (1960 film)*]

As a part of our future improvements for this crucial component of the trailer creation process, we are going to switch an algorithm presented in PySceneDetect [Castellano, 2018] to ridiculously fast SBD method, which uses Fully Convolutional Networks [Long, Shelhamer, and Darrell, 2014], presented in [Gygli, 2018]. That would give us an ability to achieve *120x real-time speed* and to detect shot changes much more accurately than all the previous SBD methods. That runtime boost is particularly valuable for us since one of the most important advantages of automatic trailer generation methods over real professional trailer editors should be the cost time of trailer creation.

4.2 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 change 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 simply need to add a custom feature, which is responsible for detecting explosions (can be done either with the video or audio feature). A lot of time was spent on the in-depth discovery of both visual and aural features, but during their validation, some of them were removed. All three types of features (visual, audio short-term, and audio mid-term) can be observed in Table 4.1.

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

4.2.1 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, *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 *content value* parameter (mean between Hue, Saturation, and Lightness) to this group of features, as it takes the most significant role in the 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. Creation of these features was achieved by leveraging the capabilities of Faster R-CNN [Ren et al., 2015], pretrained on MS COCO dataset [Lin et al., 2014]. 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 stuff in their natural context. From the extracted information about objects on the frame, we construct 6 features, that 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 actions, which keeps viewers attention on the screen. Another group was formed under the assumption that close-up shots are attractive to view. The complete list of visual properties that we capture at the frame level can be inspected in Figures 4.3 and 4.4.

FIGURE 4.3: An example of all frame-level features on single frame with single person.

FIGURE 4.4: An example of all frame-level features on single frame with multiple people and multiple other objects.

4.2.2 Audio short-term and mid-term features

In most of the cases, the most salient audio parts are accompanied by outstanding visual scenes and vice versa. Therefore, audio features are no less important than the visual ones. For our case, we have applied a set of audio features previously introduced in [Giannakopoulos et al., 2014]. All audio features were retrieved by exploiting the potential of the open-source library for audio signal analysis named *pyAudioAnalysis* [Giannakopoulos, 2015]. The main factor of our choice during feature selection was that these features had been used in multiple audio analysis and processing techniques because of their significant coverage of sound signal properties.

Before the feature extraction step, an audio signal is usually cut into short non-overlapping windows (frames). For the short-term feature sequences, we have used a frame size of 50 msecs of an audio signal, and 1-second window size for the mid-term, correspondingly. As a result of feature extraction, for short-term and mid-term audio signals, we get a sequence of 34-dimensional and 68-dimensional corresponding feature vectors per each frame. **Short-term features** include fundamental audio properties:

- **Zero Crossing Rate** is the rate of changes of sign of the signal divided by its duration. This property is the most important for detecting percussive sounds. It can be interpreted as a measure of the signal noisiness.
- **Energy** is calculated as the normalized by the frame length sum of the squares of input signal values. As a rule, over consecutive speech frames, short-term energy shows high variance, due to brief intervals of silence between the words. Such an audio feature could be particularly helpful for segmenting speech parts from the audio.
- **Entropy of Energy** can be explained as a measure of sudden and unexpected changes in the energy level of the input audio signal.
- **Spectral Centroid** and **Spectral Spread** are the primary domain features that are responsible for measuring the quantity of position and shape, correspondingly. The spectral centroid is used to describe the spectrum, indicate its centroid. From the perception perspective, it is densely connected with the pattern of *"brightness"* of the audio signal. Additionally, it is widely used in audio analysis applications for determining the musical timbre. Spectral spread is computed as the average spread of the spectrum, taking into an account its centroid. The primary usage of this feature is its ability to indicate the noisiness of an audio signal accurately.
- **Spectral Entropy** is calculated similarly to the entropy of energy, although in contrast to this time domain feature spectral entropy is primarily applied in the frequency domain.
- **Spectral Flux** is usually calculated as the Euclidian distance between the two normalized magnitudes of the spectra of the two consecutive frames. It can be interpreted as a measure of how quickly the power spectrum of the audio signal is changing.
- **Spectral Rolloff** is the frequency below which a specific amount *(in our case 90%)* of the spectrum magnitude distribution is concentrated. It usually plays the role of a discriminator between voice and voiceless sounds.
- **MFCCs** (The Mel-Frequency Cepstrum Coefficients) are coefficients that together form MFC (Mel-Frequency Cepstrum), where the non-linear frequency bands are distributed according to the Mel-scale. They are commonly used in a wide variety of applications: speech recognition, music information retrieval, speaker clustering, and for lots of other audio analysis purposes.
- **Chroma Vector** *(also called chromagram)* is the 12-dimensional vector of values, which represents the spectral energy of an audio signal. The main advantage of chroma features is that they catch harmonious and melodic music characteristics, at the same time being robust to mutations of timbre. In practice,

all 12 chroma values are representing the 12 equal-tempered pitch classes of western-type music. Chroma features are an essential component of the most audio and speech analysis systems.

- **Chroma Deviation** is the std *(standard deviation)* of the 12 chroma coefficients of the chroma vector.

Mid-term features play a role of 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.

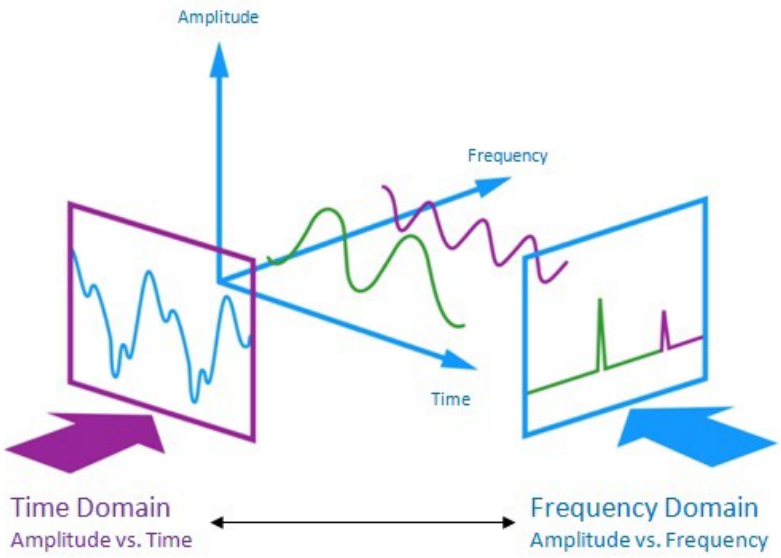

FIGURE 4.5: The structure of an audio signal. Source: [Doshi, 2018]

To sum it up, we have gathered together all the essential properties of the audio signal for both time and frequency domains (Figure 4.5), that could be further utilized for multiple purposes: from detecting speech among other sounds to determining the saliency of different parts of the audio.

4.3 Anomalous scene selection

Anomalous scenes selection is a long process containing multiple steps: anomaly detectors selection, retrieving anomaly frames for each type of features, choice of abnormal visual, 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, scene reconstruction, threshold-based anomalous scene selection. The complete pipeline of this scene selection approach can be investigated in Figure 4.6.

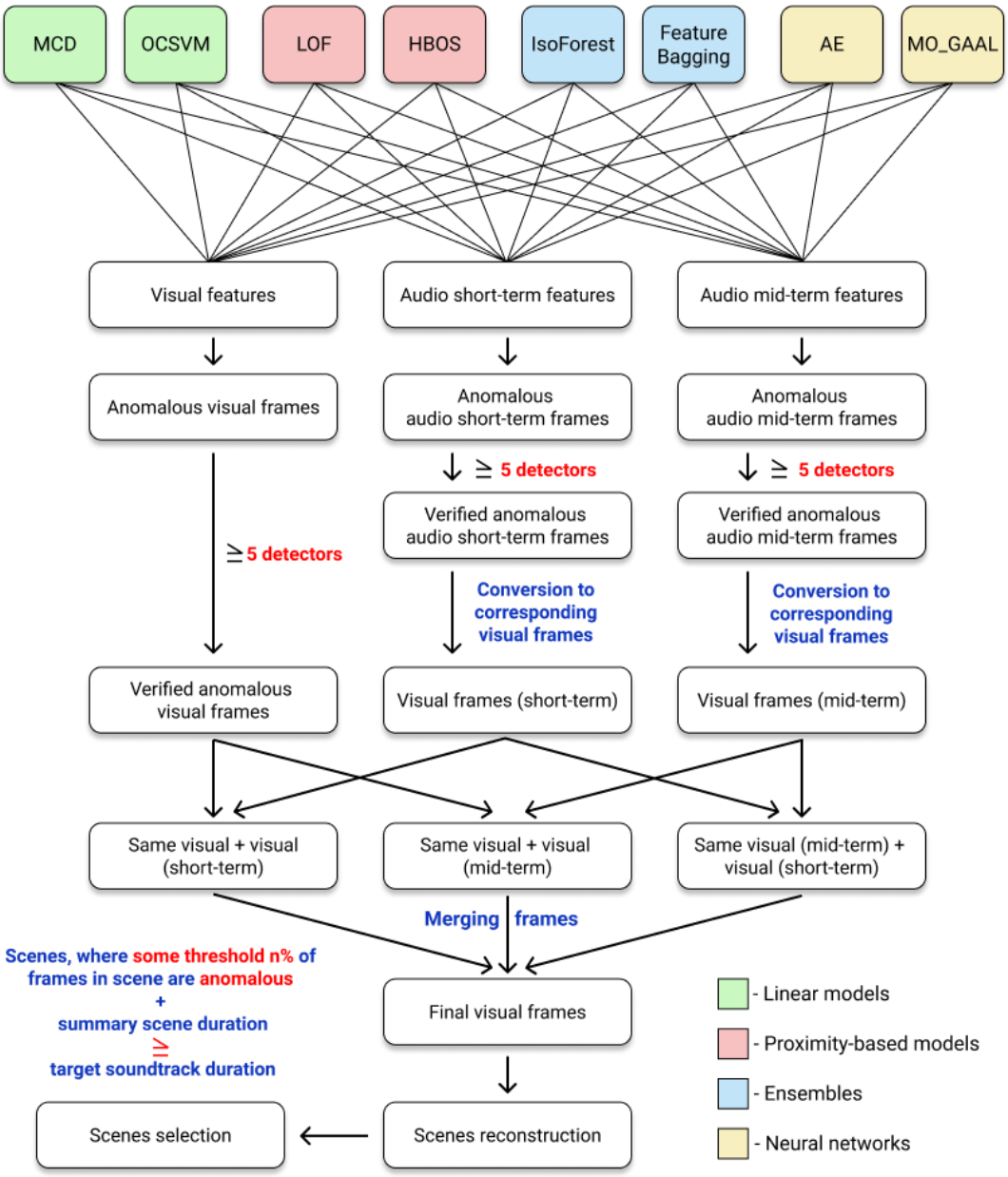

FIGURE 4.6: The detailed pipeline of anomalous scenes selection.

4.3.1 Detectors

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 detector/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 that would cover most of these cases, that could be divided into 4 groups (2 detectors per each group):

- **Linear models:**
  - **MCD** (Minimum Covariance Determinant) [Rousseeuw and Driessen, 1999] is a highly robust estimator of covariance. Being resistant to outlying observations makes the MCD very helpful in outlier detection.
  - **OCSVM** (One-Class SVM) [Manevitz and Yousef, 2001] is an unsupervised outlier detector, which is an extension of SVM (Support Vector Machine) and is regularly used for detecting novelty in the data.
- **Proximity-based models:**
  - **LOF** (Local Outlier Factor) [Breunig et al., 2000] is the proximity-based unsupervised anomaly detector, where the abnormal data samples are found by measuring their local deviation with regard to their neighbours.
  - **HBOS** (Histogram-based Outlier Score) [Goldstein and Dengel, 2012] is an efficient unsupervised, which assumes the feature independence and calculates the degree of deviation by building histograms to choose outliers.
- **Ensembles:**
  - **IsoForest** (Isolation Forest) [Liu, Ting, and Zhou, 2008] is an efficient and accurate tree ensemble method with the focus on anomaly isolation, rather than normal instances profiling. This detector is distinguishable among others, because of its capacity of handling high-volume datasets.
  - **Feature Bagging** (also called Random Subspace Method) [Ho, 1998] is an ensemble learning method that fits several base estimators on different data sub-samples and then use a certain combination method (for e.g. - averaging) in order to improve accuracy and reduce overfitting.
- **Neural networks:**
  - **AE** (fully-connected AutoEncoder) [Hinton and Zemel, 1993] is a type of neural networks for learning valuable data representations directly from data. AE also could be applied to detect outliers in the data by calculating the reconstruction errors.
  - **MO_GAAL** (Multiple-Objective Generative Adversarial Active Learning) [Liu et al., 2018] is an efficient estimator that generates the outlier candidates in order to support the classifier in defining a boundary that can adequately segregate abnormal data from the normal one.

4.3.2 Anomalous frames selection

With the selected 8 detectors, we run them separately on each of the 3 types of 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 more or equal than 5 among 8 detectors have chosen so. Having selected frames of each feature type, we needed to reduce audio short-term and mid-term frames to their corresponding visual frames taking into consideration the duration periods of each feature group frame. For clearness, let's name reduced to visual frames audio short-term and mid-term frames as visual short-term and visual mid-term, respectively. The following way of selecting the final video frames for further scenes reconstruction was decided due to multiple experiments with frames choosing algorithms. For each group of frames, we take intersection with another group, and as a result, we receive 3 new groups: visual $\cap$ visual short-term, visual $\cap$ visual mid-term and visual short-term $\cap$ visual mid-term. After that, we take an intersection between all groups and in such way form a set of anomalous final video frames. These frames served as the basis to identify trailer-worthy scenes from an input movie.

4.3.3 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 constraint for scenes selection is to select scenes with the maximum percentage of anomalous frames, while the total duration of all scenes should be not less than the length of the accompanying soundtrack. Through the visual examination of selected scenes, we have determined that these scenes with the highest number of abnormal frames are the most valid candidates for making the trailer.

4.4 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 validated a hypothesis that lots of percussive timbres (claps, snares, drums, etc.) accompany fast shots with lots of actions. 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 [Gouyon, Pachet, and Delerue, 2000], 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, 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 timeline in the original movie. Since the visual appearance of the arranged by audio feature trailers was honestly worse than the one ordered by chronology, we consequently stuck with the latter option.

Chapter 5

Evaluation and Results

Since there is no reliable metric for estimating the visual perception, either determining the level of attractiveness or aesthetics of the viewed video, we have decided to go with the subjective evaluation in different forms. In order to validate our approach, we have divided our evaluation process into two parts:

- Comparison to the original trailer
- Comparison to the leading competitors

5.1 Comparison to the original trailer

For the purpose of the comprehensive evaluation of generated with our approach trailer of *"Requiem for a dream"* vis-à-vis the original one, we have divided this part of the assessment in the following components:

- Scoring
- Visual inspection

5.1.1 Scoring

For the scoring section, likewise the authors of [Smith et al., 2017], we have asked 23 volunteers to judge these two trailers and give a score from 1 to 10 for the following questions:

- Give the overall rating for this trailer.
- How strongly this trailer arouse interest in you?
- How many spoilers does this trailer contain?
- How likely would you watch the original movie after viewing this trailer?

The results, shown in Figure 5.1 reveals that generated with **movie2trailer's** trailer demonstrate pretty competitive results even against the original trailer in terms of visual appearance and interest arousal, taking into an account that **movie2trailer** is an automatic unsupervised algorithm in less than 24 hours, and in contrast, the original *"Requiem for a dream"* trailer was created by professional editors in days or even weeks. But still based on the results, our approach needs to be improved in terms of showing too much of the movie events.

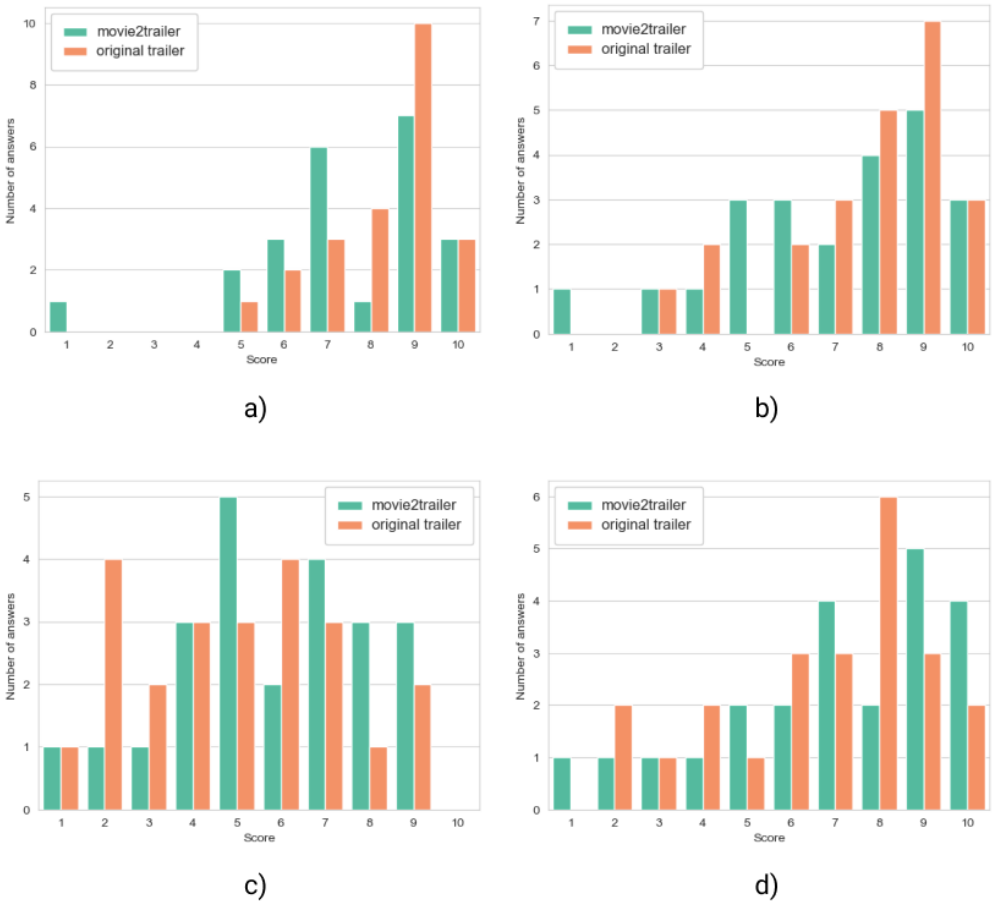

FIGURE 5.1: Comparing distributions of user ratings for 4 different questions for all 23 volunteers for the movie *"Requiem for a dream"*. Ratings range between 0 (very low) to 10 (very high). **movie2trailer** response is in green while the original response is shown in orange. Questions compared are: (a) Overall rating for the trailer, (b) How strongly trailer arouse interest in you, (c) The trailer gives too much of movie, (d) Would you watch the movie after watching this trailer.

5.1.2 Visual inspection

In Figure 5.2, we present the results of the visual inspection of our **movie2trailer** generated trailer of *"Requiem for a dream"* against the original one. For each of the trailers, there have been chosen eight scenes. Frames representing the scenes are arranged temporally (from top to bottom), with the original trailer on the left and the one generated with our approach on the right. The frames from the common scenes are connected with two-sided arrows. That is important to mention that numerous scenes from the original *"Requiem for a dream"* trailer were selected as the best candidates for the trailer generation with our AI system. The fact that we have chosen a specific set of aural and visual features shows that the scenes proposed with our scene selection method correspond to the expected result since we can explain why each of the scenes was picked. Moreover, this clearly indicates that the scenes selected for creating a movie trailer are not chosen randomly, and we can easily manipulate the outcomes of our scene selection method by tweaking the input audio and video features.

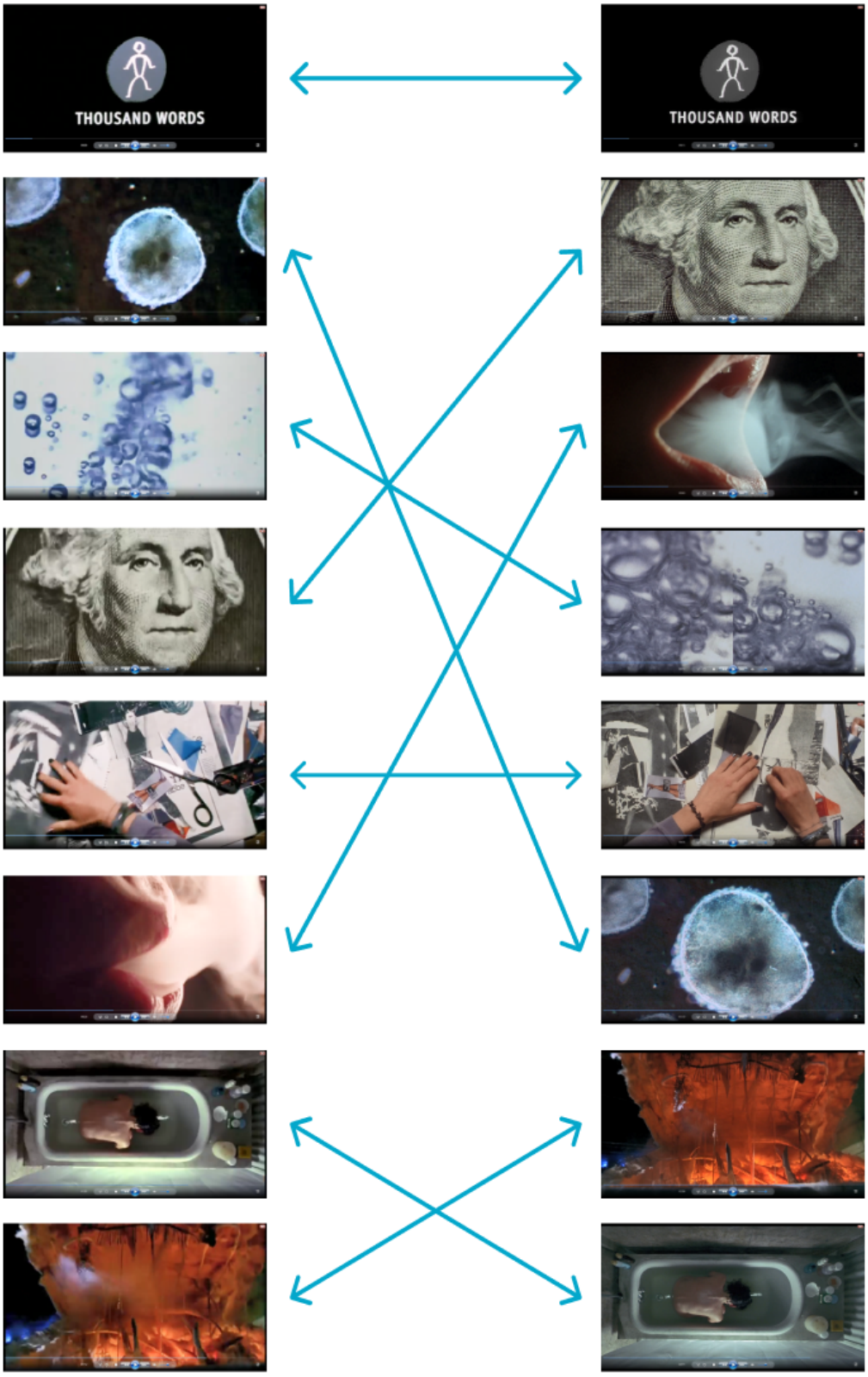

FIGURE 5.2: Selected scenes from the *"Requiem for a dream"* trailers arranged by timeline from top to bottom. The original trailer is shown on the left while the trailer generated with our approach is on the right. Arrows highlight common scenes used in both trailers.

5.2 Comparison to main competitors

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

- Trailer generation method **"V2T"** (Video2Trailer) [Irie et al., 2010];
- Commercial software for video summarization **"Muvee"**;
- **SOTA** for automatic trailer generation **"PPBVAM"** (Point Process-Based Visual Attractiveness Model) [Xu, Zhen, and Zha, 2015];
- The original official real trailers **"RT"**;
- The same real trailers without speech information **"RTwS"**.

For the fairness and objective evaluation, we have downscaled our final generated trailers to the resolution of other trailers (480x360) produced by our competitors' approaches and removed all the speech pieces from it. 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 [Irie et al., 2010] an [Xu, Zhen, and Zha, 2015], 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 3 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 [Likert, 1932]: from 1 (the lowest) to the 7 (the highest). Figure 5.3 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)"*. 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. Because all trailers generated using automatic trailer generation methods were deprived of speeches, subtitles, and special effects of montages, **"RTwS"** and **"RT"** were usually preferred more by volunteers. Since the information that these factors provide to improving visual attractiveness, we should also supplement our system with these information sources.

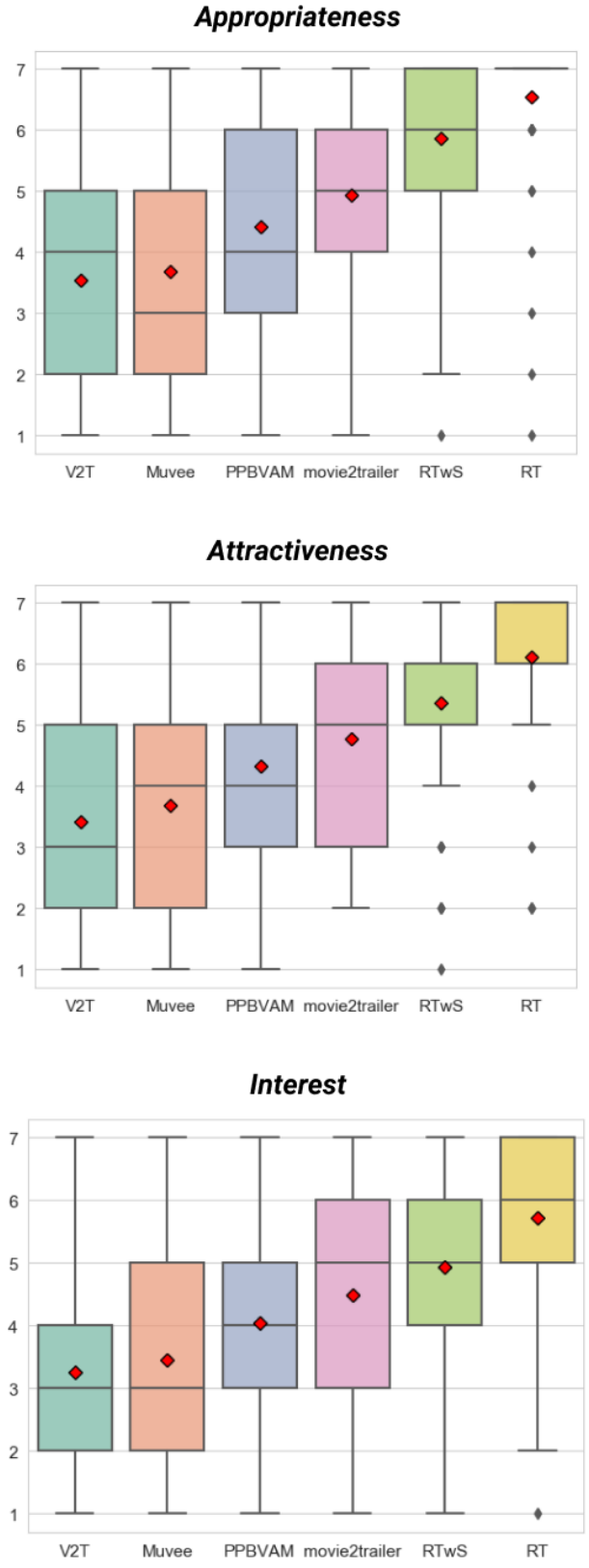

FIGURE 5.3: The box plots of scores for various methods on three questions considering Appropriateness, Attractiveness and Interest. The dark lines inside boxes are medians and red diamonds are means. Dark points outside of the whiskers are outliers.

Chapter 6

Conclusions and Future work

In this paper, we 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 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 conventional 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.

Our future works include possible improvements on all steps of the trailer creation process. Since computation cost is a critical factor for automatic trailer generation, optimizing every step would be very beneficial for us. First of all, we would replace the content-aware shot boundary detection method with the more advanced and significantly faster SBD algorithm, which uses FCN. Secondly, we would provide more extensive exploration considering the selection of visual and audio features to maximize the attractiveness. Also, we would remove or replace some of the anomaly detectors to reduce the runtime while preserving the quality of anomalous frames selection. Considering shots reordering, we plan to move to the supervised methods. As an optional but beneficial step, we want to apply synchronization of shot changes and percussive sounds in the accompanying trailer audio to enhance human perception. In the future, we are interested in augmenting our approach to utilize other information such as speeches, titles, and subtitles to create trailers with all the features that are present in the "real" trailers. We're very excited about pushing the possibilities of how AI can simulate the expertise and creativity of the professional movie editors.

Bibliography

Abdulhussain, Sadiq H. et al. (2018). "Methods and Challenges in Shot Boundary Detection: A Review". In: *Entropy* 20.4, p. 214. DOI: 10.3390/e20040214. URL: https://doi.org/10.3390/e20040214.

Ahmed, Mohiuddin, Abdun Naser Mahmood, and Jiankun Hu (2016). "A survey of network anomaly detection techniques". In: *J. Network and Computer Applications* 60, pp. 19–31. DOI: 10.1016/j.jnca.2015.11.016. URL: https://doi.org/10.1016/j.jnca.2015.11.016.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2001). "Latent Dirichlet Allocation". In: *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pp. 601–608. URL: http://papers.nips.cc/paper/2070-latent-dirichlet-allocation.

Borth, Damian et al. (2013). "Large-scale visual sentiment ontology and detectors using adjective noun pairs". In: *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pp. 223–232. DOI: 10.1145/2502081.2502282. URL: https://doi.org/10.1145/2502081.2502282.

Breathless (1960 film). http://en.wikipedia.org/w/index.php?title=Breathless%20(1960%20film)&oldid=891333089. [Online; accessed 03-May-2019].

Breunig, Markus M. et al. (2000). "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*. Pp. 93–104. DOI: 10.1145/342009.335388. URL: https://doi.org/10.1145/342009.335388.

Cahuina, Edward J. Y. Cayllahua and Guillermo Cámara Chávez (2013). "A New Method for Static Video Summarization Using Local Descriptors and Video Temporal Segmentation". In: *XXVI Conference on Graphics, Patterns and Images, SIBGRAPI 2013, Arequipa, Peru, August 5-8, 2013*, pp. 226–233. DOI: 10.1109/SIBGRAPI.2013.39. URL: https://doi.org/10.1109/SIBGRAPI.2013.39.

Castellano, Brandon (2018). *Video Scene Cut Detection and Analysis Tool*. https://github.com/Breakthrough/PySceneDetect.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). "Anomaly detection: A survey". In: *ACM Comput. Surv.* 41.3, 15:1–15:58. DOI: 10.1145/1541880.1541882. URL: https://doi.org/10.1145/1541880.1541882.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1, pp. 1–38.

Dijkstra, E. W. (1959). "A Note on Two Problems in Connexion with Graphs". In: *Numerische Mathematik* 1, pp. 269–271.

Doshi, Sanket (2018). "Music Feature Extraction in Python". In: *Towards Data Science*. URL: https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d.

Eyben, Florian, Martin Wöllmer, and Björn W. Schuller (2009). "OpenEAR - Introducing the munich open-source emotion and affect recognition toolkit". In: *Affective Computing and Intelligent Interaction, Third International Conference and Workshops, ACII 2009, Amsterdam, The Netherlands, September 10-12, 2009, Proceedings*, pp. 1–6. DOI: 10.1109/ACII.2009.5349350. URL: https://doi.org/10.1109/ACII.2009.5349350.

— (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pp. 1459–1462. DOI: 10.1145/1873951.1874246. URL: https://doi.org/10.1145/1873951.1874246.

Eyben, Florian et al. (2013). "Recent developments in openSMILE, the munich open-source multimedia feature extractor". In: *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pp. 835–838. DOI: 10.1145/2502081.2502224. URL: https://doi.org/10.1145/2502081.2502224.

Frey, B.J. and D. Dueck (2007). "Clustering by passing messages between data points". In: *science* 315.5814, pp. 972–976.

Giannakopoulos, Theodoros (2015). "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis". In: *PloS one* 10.12.

Giannakopoulos, Theodoros et al. (2014). "Realtime depression estimation using mid-term audio features". In: *Proceedings of the 3rd International Workshop on Artificial Intelligence and Assistive Medicine co-located with the 21th European Conference on Artificial Intelligence (ECAI 2014), Prague, Czech Republic, August 18, 2014.* Pp. 41–45. URL: http://ceur-ws.org/Vol-1213/paper9.pdf.

Goldstein, Markus and Andreas Dengel (2012). "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm". In: *KI-2012: Poster and Demo Track*, pp. 59–63.

Goldstein M, Uchida S (2016). "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data". In: *PLoS ONE*. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152173.

Goodfellow, Ian J. et al. (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.

Goto, Masataka (2006). "A chorus section detection method for musical audio signals and its application to a music listening station". In: *IEEE Trans. Audio, Speech & Language Processing* 14.5, pp. 1783–1794. DOI: 10.1109/TSA.2005.863204. URL: https://doi.org/10.1109/TSA.2005.863204.

Gouyon, Fabien, François Pachet, Olivier Delerue, et al. (2000). "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*.

Gygli, Michael (2018). "Ridiculously Fast Shot Boundary Detection with Fully Convolutional Neural Networks". In: *2018 International Conference on Content-Based Multimedia Indexing, CBMI 2018, La Rochelle, France, September 4-6, 2018*, pp. 1–4. DOI: 10.1109/CBMI.2018.8516556. URL: https://doi.org/10.1109/CBMI.2018.8516556.

Hauptmann, Alexander G. et al. (2007). "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*, pp. 20–24. DOI: 10.1145/1290031.1290034. URL: https://doi.org/10.1145/1290031.1290034.

Hinton, Geoffrey E. and Richard S. Zemel (1993). "Autoencoders, Minimum Description Length and Helmholtz Free Energy". In: *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pp. 3–10. URL: http://papers.nips.cc/paper/798-autoencoders-minimum-description-length-and-helmholtz-free-energy.

Ho, Tin Kam (1998). "The Random Subspace Method for Constructing Decision Forests". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 20.8, pp. 832–844. DOI: 10.1109/34.709601. URL: https://doi.org/10.1109/34.709601.

Hou, Xiaodi, Jonathan Harel, and Christof Koch (2012). "Image Signature: Highlighting Sparse Salient Regions". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 34.1, pp. 194–201. DOI: 10.1109/TPAMI.2011.146. URL: https://doi.org/10.1109/TPAMI.2011.146.

Huang, Chengqiang (2018). "Featured anomaly detection methods and applications". PhD thesis. University of Exeter, Devon, UK. URL: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.761755.

Irie, Go et al. (2010). "Automatic trailer generation". In: *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pp. 839–842. DOI: 10.1145/1873951.1874092. URL: https://doi.org/10.1145/1873951.1874092.

Isham, Valerie and Mark Westcott (1979). "A self-correcting point process". In: *Stochastic Processes and their Applications* 8 (3), pp. 335–347.

Itti, Laurent and Pierre Baldi (2005). "Bayesian Surprise Attracts Human Attention". In: *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pp. 547–554. URL: http://papers.nips.cc/paper/2822-bayesian-surprise-attracts-human-attention.

Ji, Zhong et al. (2017). "Video Summarization with Attention-Based Encoder-Decoder Networks". In: *CoRR* abs/1708.09545. arXiv: 1708.09545. URL: http://arxiv.org/abs/1708.09545.

Kibish, Sergey (2018). "A note about finding anomalies". In: *Towards Data Science*. URL: https://towardsdatascience.com/a-note-about-finding-anomalies-f9cedee38f0b.

Kuwano, Hidetaka et al. (2000). "Telop-on-Demand: Video Structuring and Retrieval based on Text Recognition". In: *2000 IEEE International Conference on Multimedia and Expo, ICME 2000, New York, NY, USA, July 30 - August 2, 2000*, pp. 759–762. DOI: 10.1109/ICME.2000.871472. URL: https://doi.org/10.1109/ICME.2000.871472.

Lienhart, Rainer (1999). "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*, pp. 290–301. DOI: 10.1117/12.333848. URL: https://doi.org/10.1117/12.333848.

Likert, Rensis (1932). "A technique for the measurement of attitudes." In: *Archives of psychology*.

Lin, Tsung-Yi et al. (2014). "Microsoft COCO: Common Objects in Context". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48. URL: https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou (2008). "Isolation Forest". In: *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pp. 413–422. DOI: 10.1109/ICDM.2008.17. URL: https://doi.org/10.1109/ICDM.2008.17.

Liu, Yezheng et al. (2018). "Generative Adversarial Active Learning for Unsupervised Outlier Detection". In: *CoRR* abs/1809.10816. arXiv: 1809.10816. URL: http://arxiv.org/abs/1809.10816.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell (2014). "Fully Convolutional Networks for Semantic Segmentation". In: *CoRR* abs/1411.4038. arXiv: 1411.4038. URL: http://arxiv.org/abs/1411.4038.

Ma, Yu-Fei et al. (2002). "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*. Pp. 533–542. DOI: 10.1145/641007.641116. URL: https://doi.org/10.1145/641007.641116.

Mahasseni, Behrooz, Michael Lam, and Sinisa Todorovic (2017). "Unsupervised Video Summarization with Adversarial LSTM Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2982–2991. DOI: 10.1109/CVPR.2017.318. URL: https://doi.org/10.1109/CVPR.2017.318.

Manevitz, Larry M. and Malik Yousef (2001). "One-Class SVMs for Document Classification". In: *Journal of Machine Learning Research* 2, pp. 139–154. URL: http://jmlr.org/papers/v2/manevitz01a.html.

Minami, Kenichi et al. (1998). "Video Handling with Music and Speech Detection". In: *IEEE MultiMedia* 5.3, pp. 17–25. DOI: 10.1109/93.713301. URL: https://doi.org/10.1109/93.713301.

Ogata, Y. and D. Vere-Jones (1984). "Inference for earthquake models: A self-correcting model". In: *Stochastic Processes and their Applications* 17 (2), pp. 337–347.

Plutchik, Robert (2001). "The nature of emotions". In: *American Scientist* 89.4, pp. 344–350.

Ren, Shaoqing et al. (2015). "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*, pp. 91–99. URL: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.

Rousseeuw, Peter J. and Katrien van Driessen (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator". In: *Technometrics* 41.3, pp. 212–223. DOI: 10.1080/00401706.1999.10485670. URL: https://doi.org/10.1080/00401706.1999.10485670.

Smith, John R. et al. (2017). "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*, pp. 1799–1808. DOI: 10.1145/3123266.3127906. URL: https://doi.org/10.1145/3123266.3127906.

Wei, Huawei et al. (2018). "Video Summarization via Semantic Attended Networks". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 216–223. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16581.

Wu, Jiaxin et al. (2018). "Foveated convolutional neural networks for video summarization". In: *Multimedia Tools Appl.* 77.22, pp. 29245–29267. DOI: 10.1007/s11042-018-5953-1. URL: https://doi.org/10.1007/s11042-018-5953-1.

Xu, Hongteng, Yi Zhen, and Hongyuan Zha (2015). "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, pp. 2198–2204. URL: http://ijcai.org/Abstract/15/311.

YouTube. https://www.youtube.com/watch?v=OLDhra9g8Lc.

Yuan, Jinhui et al. (2007). "A Formal Study of Shot Boundary Detection". In: *IEEE Trans. Circuits Syst. Video Techn.* 17.2, pp. 168–186. DOI: 10.1109/TCSVT.2006.888023. URL: https://doi.org/10.1109/TCSVT.2006.888023.

Zhou, Bolei et al. (2014). "Learning Deep Features for Scene Recognition using Places Database". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 487–495. URL: http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.

Zhuang, Xinhua et al. (1996). "Gaussian mixture density modeling, decomposition, and applications". In: *IEEE Trans. Image Processing* 5.9, pp. 1293–1302. DOI: 10.1109/83.535841. URL: https://doi.org/10.1109/83.535841.