BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 6 (2021) 1–12

BRILL OPEN

brill.com/rdj

A Data Set for US Horror Film Trailers

Arts and Media

Nick Redfern
Independent researcher, Leeds, UK
nickredfernres@outlook.com

Abstract

This article presents a new data set comprising audio, colour, motion, and shot length data of trailers for the fifty highest grossing horror films at the US box office from 2011 to 2015. This data set is one of the few available for computational film analysis that includes data on multiple elements of film style and is the only existing data set for motion picture trailers suitable for formal analyses. Data is stored in csv files available under a Creative Commons Attribution 4.0 International license on Zenodo: www.doi.org/10.5281/zenodo.4479068.

Keywords

computational film analysis – horror cinema – film trailers – sound – colour – editing – motion

Online publication date: 09-11-2021

– Related data set "US Horror Trailers Data Set" with DOI www.doi.org/10.5281/zenodo.4479068 in repository "Zenodo"

1. Introduction

David Bordwell describes film style as "a film's systematic and significant use of the techniques of the medium … Style is, minimally, the texture of the film's images and sounds, the result of choices made by filmmaker(s) in particular historical circumstances" (1997, p. 4). Film style can be divided into four broad

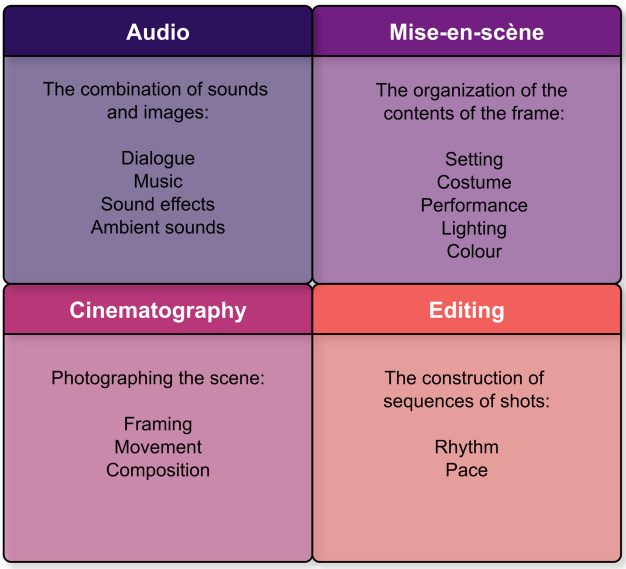

FIGURE 1 The four elements of film style

categories: mise-en-scène, cinematography, editing, and audio (see Figure 1). Our purpose in analysing the style of a film is "to explain why an individual motion picture is the way it is: why it has the elements of style it does and why they stand in the relations that they do" (Carroll, 2009, p. 268). Computational film analysis lies squarely within the digital humanities, and aims to answer questions about film style framed from within the tradition of the humanities using computational methods (Heftberger, 2018; Redfern 2020a), employing the methods and tools of statistics, data science, information visualisation, and computer science in order to understand the formal properties of the cinema. Distributing films as a corpus is not possible due to copyright restrictions. Therefore, making data derived from films under analysis available along with a description of the methods by which that data was collected and processed is a key part of computational film analysis, promoting collaboration, innovation in approaches to analysing film style, reproducibility, and transparency.

None of the extant movie trailer data sets available were designed to answer questions about film style and are not suitable for computational film analysis for two reasons. First, some data sets contain no data about or derived from movie trailers: although both the MovieLens (Abu-El-Haija et al., 2018) and MovieNet (Huang et al., 2020) data sets include useful data on other aspects of the cinema, the only information relating to trailers available are the YouTube IDs of the films listed in the main part of the data sets. Second,

data sets containing aesthetic information about trailers such as the MMTF-14K data set (Deldjoo et al., 2018) and the LMTD (Wehrmann & Barros, 2017) are designed for classifying trailers for movie recommendation systems, and so the aesthetic data they make available has been reduced to feature vectors suitable for processing by machine learning algorithms. Furthermore, those features only represent some computable aspects of film style, with other key aspects not included: the aesthetic dimensions of the MMTF-14K data set, for example, relate to sound, colour, texture, and object recognition and does not include any cinematic (i.e., motion or editing) data.

In this article I present a data set with data on multiple elements and comprising audio, colour, movement, and editing data for trailers for the fifty highest grossing horror films at the US box office from 2011 to 2015, describing the main features of the data set and how it was created. This data set is one of the few available for computational film analysis that includes data on multiple elements of film style and is the only data set available for computational film analysis of motion picture trailers.

2. Style in Motion Picture Trailers

Trailers provide audiences with an opportunity to sample films prior to release and are one of the most effective forms of film advertising (Karray & Debernitz, 2017). They are a part of the experience of cinema, both in a theatre when screened prior to a film and as part of a broader cinematic culture that includes a film's paratexts. Trailers exist as hybrid cinematic forms that must fulfil a range of functions in a brief time frame of (typically) between 90 and 150 seconds. As short films they must create a sufficiently interesting experience to inspire the viewer to purchase a ticket for the film promoted, providing key information about narrative, characters, and the thrills on offer, whilst simultaneously communicating key marketing information, such as the release date, social media information, and relationships to other films in a franchise or by the same producers. To achieve these objectives, trailers present a highly compressed monomaniacal version of a film that reduces narrative complexity and intensifies the visual and sonic elements of film style.

Trailers are produced by companies specialising in turning a feature length movie into a short film with an understanding of marketing and are released before a film is completed. This means that while they inherit some of their formal features from their reference film, such as the colour palette and the onscreen motion in shots selected for inclusion, those elements are reworked to meet the varying demands of a trailer. Other elements are introduced separately

from the mode of production of the advertised film. Trailer soundtracks are produced up to six months in advance of a film's release by sound designers using available sounds from a film in combination with sounds from effects libraries, so that many of the sounds heard in a trailer may not, in fact, appear in the movie they are used to promote (Redfern, 2020b). The editing of trailers constructs sequences characterised by a rhythm and a pace that differs sharply from the film from which the shots are taken, with the extensive use of jump cuts and montages to create connections between scenes and characters that exist in only in the trailer.

Understanding how the style of horror film trailers allows them to fulfil their functions as both entertaining short films and as promotional films therefore requires an approach to the analysis of style in the cinema that considers the different elements of style in relation to one another rather than individually. This requirement is the motivation for this data set – to put the different elements of style in relation to one another to see how they function as a dynamic system, an approach Eisenstein ([1943] 1986) described as *the synaesthetic mode*. It is not a data set for classifying texts by machine learning or viewing at a distance. It is a data set for formal analysis of a class of motion pictures (trailers for horror films) and is intended to meet the demands of close film analysis – to explain why the elements of a motion picture are arranged in the way they are – and to identify trends across a sample of similar texts. What is at stake in this data set is not that the motion pictures covered are trailers for horror films but *how* they create a frightening experience for the viewer to entice them into viewing for a film and the aesthetic norms used by trailer producers within this genre (Hanich, 2010).

3. Methods

3.1. *Selection of Trailers*

As a generic label I use *horror* as defined in Redfern (2008) to refer broadly to

> films evoking responses of fear, terror, disgust, shock, suspense, and/or horror in the viewer through the presentation of a prototypical narrative in which a character (or characters) is confronted with a hypernatural antagonist producing autonomic responses that are transformed into telic emotional states that form the basis for (simulated) motor actions leading to the destruction of the antagonist.

Horror is a broad top-level category on www.the-numbers.com and is broken down into a set of creative types which are represented to greater and lesser degrees across the sample, including supernatural horror, natural horror, action horror, found-footage horror, gothic horror, science fiction horror, horror comedy, psychological horror, body horror, and slasher films.

The data set includes trailers for the top 50 grossing US horror films from 2011 to 2015 based on inflation-adjusted box office data from www.the-numbers.com, with one trailer per film included. Horror films outside the top 50 for this period tend to be films with limited releases or non-English language films, with much lower grosses. By focussing on the high grossing films, the films in the data set are consistent in terms of their release as US films on general release. The sample only includes trailers distributed as suitable for "appropriate audiences" (known as *green band trailers*), where "appropriate" refers to the context of viewing the trailer cinematically in relation to the main feature for which an audience has purchased a ticket.

3.2. *Pre-processing Video Files*

Trailers were downloaded from YouTube as mp4 files. Prior to data collection, the trailers were edited using DaVinci Resolve (v.16.2.7.010) to remove MPAA rating tag screens along with any promotional materials for the YouTube channel added to the top or tail of a trailer. The trailers were also cropped to remove letter-box blanking. Figure 2 describes the pre-processing workflow. These steps are essential to ensure data quality by removing features from the raw mp4 files that would otherwise contaminate the data, but impose a significant cost in terms of the labour required to produce the data set, as the video files must be edited by hand necessarily reducing the size of the data set but resulting in a gain in data quality.

3.3. *Audio Data Collection*

Following pre-processing, the soundtrack for each trailer was exported from DaVinci Resolve as a stereo 16-bit linear PCM wave at a sampling rate of 48 kHz. The wave file of each trailer soundtrack was then processed in R (v.4.0.3; R Core Team, 2020) using the packages tuneR (v. 1.3.3; Ligges et al., 2018) and seewave (v. 2.1.5; Sueur et al., 2008).

The wave file was loaded into R using tuneR::readWave() and converted to a mono wave object by averaging the amplitude of both channels of the stereo wave file using tuneR::mono(), with which = "both". The time contour of the normalized aggregated power envelope was calculated using seewave::acoustat(), with Hanning windows of 2048 samples in length overlapped by 50%. The same settings were used for every trailer and this workflow is visualised

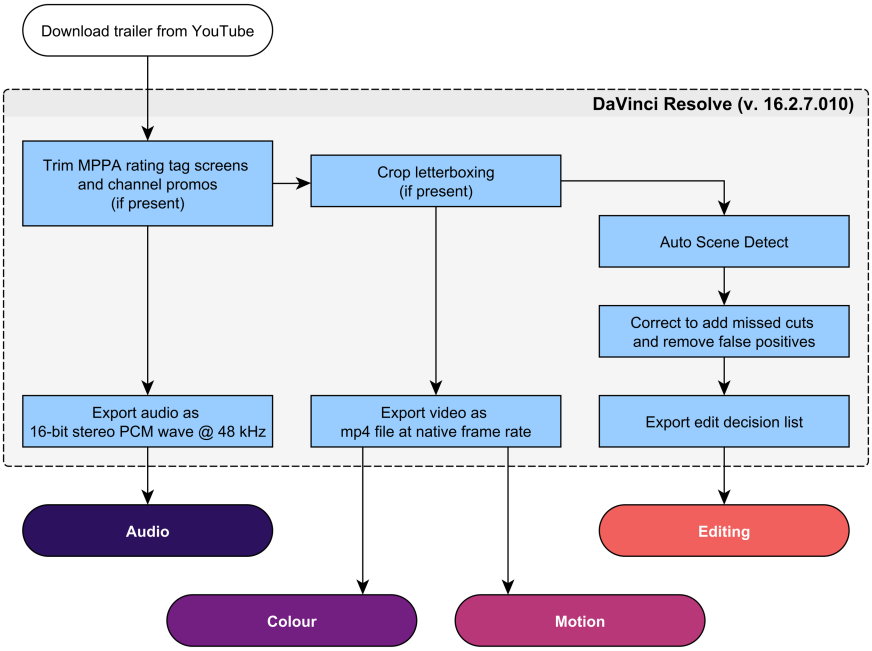

FIGURE 2 Pre-processing workflow for US horror trailers

in Figure 3. seewave::acoustat() calculates the short-time Fourier transform (STFT) of a mono wave object to produce a time-frequency matrix and sums the amplitude values of each column of the matrix to produce an aggregated power envelope that is normalized to a unit area and treated as an amplitude probability mass function. Redfern (n.d.) has a detailed discussion of the application of this method to film sound.

The trailers covered by this data set are all subject to copyright and exporting the wave object after the first step in Figure 3 is to duplicate the soundtrack as a data file. At this stage, there is no distinction between data and text: a data file containing the amplitude values comprising the wave form of a trailer's soundtrack could be converted to an S4 Wave-class object and exported from R as an audio file, thereby reproducing the soundtrack in its original form. By calculating the normalized aggregated power envelope from the STFT of the wave form, I am able to include information about a soundtrack in the data set that is meaningful and suitable for analysis while also being a transformed version of the soundtrack for distribution. The data made available in the data set is *not* the text; it is a representation of the soundtrack that enables the user to identify and analyse its key features, including the structure at different scales, the temporal evolution, the presence of affective sound events, and other features (Redfern, 2020c).

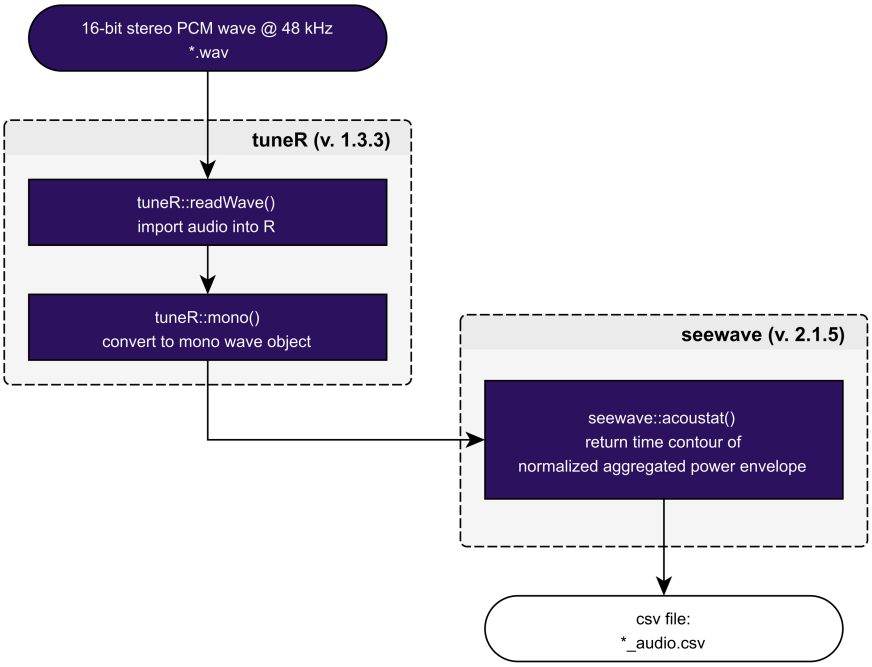

FIGURE 3 Audio data processing workflow for US horror trailers

3.4. *Colour Data Collection*

Colour data in the form of the mean RGB values for every frame of a trailer was extracted using a version of a MATLAB script by Tommaso Buonocore adapted to run in the open-source high-level programming language Octave (v. 6.1.0; Eaton et al., 2020). The original MATLAB script can be accessed via Buonocore's GitHub repository at www.github.com/detsutut/chroma (for a demonstration, see Buonocore, 2019). The adapted Octave version can be accessed via GitHub at www.github.com/DrNickRedfern/VideoProcessingOctave. Both scripts inspect a video source frame by frame and return the average colour of each frame as an RGB triplet in sRGB colour space (Figure 4).

3.5. *Motion Data Collection*

Motion data was captured with FlowAnalyzer, a Python module using the OpenCV library for computer vision that extracts motion vectors by optical flow analysis (Barbosa & Vatikiotis-Bateson, 2013, 2016). Using the complete frame as a region of interest, FlowAnalyzer compares pixel intensities between consecutive video frames and calculates the magnitude and direction of motion for each pixel from one frame to the next with a time step between consecutive

r	g	b
36.145	37.405	35.067
36.404	37.489	35.365
36.069	37.341	34.980
36.463	37.477	35.469

FIGURE 4 Collecting average RGB values from frames of the trailer for "Insidious: Chapter 2"

frames equal to $1/f$, where f is the native frame rate of the trailer. The resulting magnitude and direction vectors are summed to return a set of scalar values of optical flow per frame. A limitation of using optical flow analysis is that the data describes the overall magnitude and direction of motion in a frame and does not distinguish between different types of motion, such as motion within the frame and camera movement.

3.6. *Shot Length Data Collection*

Shot length data for each trailer was collected by searching for edits using the auto scene detect function in DaVinci Resolve before correction by hand to add any missed cuts and remove any false positives. Gradual transitions (such as dissolves, fades, etc.) are marked as cuts at the approximate mid-point between two shots. The edit decision list of timings in hh:mm:ss+frames for a trailer was then exported and converted to timings in seconds at the native frame rate of the trailer using a Microsoft Excel spreadsheet. As a common stylistic feature of contemporary horror film trailers is to include rapidly edited sequences comprising shots of only a single frame in duration, all shot lengths are given to two decimal places.

4. Data

- **US Horror Trailers Data Set deposited at Zenodo – DOI:**www.doi.org/10.5281/zenodo.4479068
- **Temporal coverage:** 2011–2015

All data in the data set is stored in csv files.

The file US_Horror_Trailers_Sample_Summary.csv summarises the video file of each trailer in the sample from which data is collected and includes the following information:

- title: the title of the film promoted as it is used in the data set
- url: the URL of the trailer on YouTube (all URLs were correct as of 27 January 2021)
- width, height: the dimensions of the video file after pre-processing
- frames: the number of frames after pre-processing
- framerate: the native frame rate (23.976/24/25/29.97/30 frames per second)
- duration: the running time of a trailer after pre-processing in seconds

The data for each element of film style are collected into zip files:

- US_Horror_Trailers_Audio_Data.zip
- US_Horror_Trailers_Colour_Data.zip
- US_Horror_Trailers_Motion_Data.zip
- US_Horror_Trailers_SL_Data.zip

Table 1 lists the variables for each element.

A standard way of representing film titles is used throughout the data set, with punctuation removed and spaces replaced by underscores: for example, *Insidious: Chapter 2* is represented as Insidious_Chapter_2.

A standard naming convention is also used for the csv files for each element of style:

- audio data is stored in files with the naming format *_audio.csv
- RGB colour data is stored in files with the format *_rgb.csv
- motion data is stored in files with the format *_motion.csv

Where * is the title of the film in its standardised form. For example, the audio data for the *Insidious: Chapter 2* trailer is included in the csv file Insidious_Chapter_2_audio.csv, its RGB colour data in the file Insidious_Chapter_2_rgb.csv, and its motion data in Insidious_Chapter_2_motion.csv.

As the shot length data for all fifty trailers is stored in a single csv file, each column uses the standard representation of the title of the promoted film.

TABLE 1 Elements of film style and variables in the US horror trailers data set

csv naming convention	Variables
Audio	
US_Horror_Trailers_Audio_Data.zip	
One file per trailer: *_audio.csv	– time: the time codes of the individual time-spectra in the STFT in steps of 0.0213 seconds. – contour: the time contour of the normalized aggregated power envelope
Colour	
US_Horror_Trailers_Colour_Data.zip	
One file per trailer: *_rgb.csv	– Three time-ordered columns (from frame 1 to frame *n*) containing the average colour of each frame as an RGB triplet.
Motion	
US_Horror_Trailers_Motion_Data.zip	
One file per trailer: *_motion.csv	– time: the onset of a frame in seconds from the beginning of the trailer at its native frame rate (*f*), increasing by a step of 1/*f*. – frame_mag: the overall magnitude of pixel displacement in a frame step – frame_x: the horizontal direction of pixel displacement – frame_x_mag: horizontal magnitude of motion – frame_y: the vertical direction of pixel displacement – frame_y_mag: vertical magnitude of motion
Shot length data	
US_Horror_Trailers_SL_Data.zip	
US_Horror_Trailers_SL_Data.csv: one column per trailer	– shot length data in seconds, data is time ordered from shot 1 to shot *n*

5. Online Tutorial

An online tutorial illustrating some methods by which this data can be visualised using R is available at https://rpubs.com/nr62_rp33/visualizing_trailers.

References

Abu-El-Haija, S., Joonseok, L., Harper, M., & Konstan, J. (2018). MovieLens 20M YouTube Trailers Dataset. www.grouplens.org/datasets/movielens/20m-youtube/.

Barbosa, A. V., & Vatikiotis-Bateson, E. (2013). *FlowAnalyzer*. www.cefala.org/FlowAnalyzer/.

Barbosa, A. V., & Vatikiotis-Bateson, E. (2016). *FlowAnalyzer: measuring behavioral motion from video*. www.cefala.org/~adriano/pubs/pdf_files/Adriano%20Vilela%20Barbosa%20and%20Eric%20Vatikiotis-Bateson%20-%202016%20-%20FlowAnalyzer%20Measuring%20behavioral%20motion%20from%20vid.pdf.

Bordwell, D. (1997). *On the history of film style*. Harvard University Press.

Buonocore, T. (2019). *Exploring chromatic storytelling in movies with R*. www.towardsdatascience.com/exploring-chromatic-storytelling-with-r-part-1-8e9ddf8d4187.

Carroll, N. (2009). Style. In P. Livingstone & C. Plantinga (Eds.), *The Routledge companion to philosophy and film* (pp. 268–278). Routledge.

Deldjoo, Y., Constantin, M. G., Ionescu, B., Schedl, M., & Cremonesi, P. (2018, June). MMTF-14K: A multifaceted movie trailer feature dataset for recommendation and retrieval. *MMSys '18: Proceedings of the 9th ACM Multimedia Systems Conference* (pp. 450–455). www.doi.org/10.1145/3204949.3208141.

Eaton, J. W., Bateman, D., Hauberg, S., & Wehbring, R. (2020). *GNU Octave version 6.1.0 manual: a high-level interactive language for numerical computations*. www.gnu.org/software/octave/doc/v6.1.0/.

Eisenstein, S. M. ([1943] 1986). *The film sense* (J. Leyda, Trans.) Faber and Faber.

Hanich, J. (2010). *Cinematic emotion in horror films and thrillers: The aesthetic paradox of pleasurable fear*. Routledge.

Heftberger, A. (2018). *Digital humanities and film studies: Visualising Dziga Vertov's work*. Springer International Publishing.

Huang, Q., Xiong, Y., Rao, A., Wang, J., & Lin, D. (2020). Movienet: A holistic dataset for movie understanding. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (Vol. 12349, pp. 709–727). Springer International Publishing. www.doi.org/10.1007/978-3-030-58548-8_41.

Karray, S., & Debernitz, L. (2017). The effectiveness of movie trailer advertising. *International Journal of Advertising*, *36*(2), 368–392. www.doi.org/10.1080/02650487.2015.1090521.

Ligges, U., Krey, S., Mersmann, O., & Schnackenberg, S. (2018). *tuneR: Analysis of Music and Speech, Version 1.3.3*. https://CRAN.R-project.org/package=tuneR.

R Core Team (2020). *R: A language and environment for statistical computing, Version 4.0.3. R Foundation for Statistical Computing, Vienna, Austria*. www.R-project.org/.

Redfern, N. (n.d.). *An Introduction to Computational Analysis of Film Audio in R*. Retrieved January 27, 2021, from www.academia.edu/43198172/An_introduction_to_computational_analysis_of_film_audio_in_R.

Redfern, N. (2008). *Emotion, genre, and the Hollywood paranoid film, New Nightmares: Issues and Themes in Contemporary Horror Cinema, Manchester Metropolitan University, 3–4 April 2008*. www.academia.edu/2203600/Emotion_Genre_and_the_Hollywood_Paranoid_Film.

Redfern, N. (2020a). What is computational film analysis? www.computationalfilmanalysis.wordpress.com/2020/07/06/what-is-cfa/.

Redfern, N. (2020b). Sound in horror film trailers. *Music, Sound, and the Moving Image*, *14*(1), 48–71. www.doi.org/10.3828/msmi.2020.4.

Redfern, N. (2020c). Quantitative analysis of sound in a short horror film. *Humanities Bulletin*, *2*(3), 246–257. www.journals.lapub.co.uk/index.php/HB/article/view/1682/1350.

Sueur J., Aubin T., & Simonis, C. (2008). seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*(2), 213–226. www.doi.org/10.1080/09524622.2008.9753600.

Wehrmann, J., & Barros, R. C. (2017). Movie genre classification: A multi-label approach based on convolutions through time. *Applied Soft Computing*, 61, 973–982. www.doi.org/10.1016/j.asoc.2017.08.029.