Understanding videos at scale: How to extract insights for business research

Schwenzow, Jasper; Hartmann, Jochen; Schikowsky, Amos; Heitmann, Mark

Published in:
Journal of Business Research

DOI:
10.1016/j.jbusres.2020.09.059

Publication date:
2021

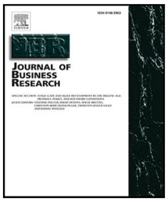
Citation for published version (APA):
Schwenzow, J., Hartmann, J., Schikowsky, A., & Heitmann, M. (2021). Understanding videos at scale: How to extract insights for business research. *Journal of Business Research*, *123*, 367-379. https://doi.org/10.1016/j.jbusres.2020.09.059

Journal of Business Research 123 (2021) 367–379

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres



Understanding videos at scale: How to extract insights for business research☆

Jasper Schwenzow*, Jochen Hartmann, Amos Schikowsky, Mark Heitmann

University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany

ARTICLE INFO

Keywords:
Video mining
Image analysis
Object detection
Deep learning
Computer vision
Unstructured data

ABSTRACT

Video content has become a major component of total internet traffic. Growing bandwidth and computational power conspire with an increasing number of video editing tools, smartphones, and online platforms that have facilitated video production, distribution, and consumption by businesses and consumers alike. This makes video content relevant across business research disciplines. However, analyzing videos can be a cumbersome manual task. Automated techniques are scattered across technical publications and are often not directly accessible to business researchers. This article synthesizes the current state of the art and provides a consolidated tool to efficiently extract 109 video-based variables, requiring no programming knowledge. The variables include structural video characteristics such as colorfulness as well as advanced content-related features such as scene cuts or human face detection. The authors discuss the research potential of video mining, the types of video features of likely interest, and illustrate application using a practical example.

1. Introduction

With the emergence of television, video has become a major form of communication between companies and consumers and has replaced former media to a large extent (Kretschmer & Peukert, 2020). As videos appeal to both the visual and acoustic senses, they tend to be more engaging than other forms of content (e.g., Choi & Johnson, 2005; Dhaoui & Webster, 2020). This growing importance is vastly amplified with increasing bandwidth and the widespread availability of video recording and editing tools, making video a medium of communication for any corporate user as well as the general population. According to industry estimates, by 2022, 82% of all internet traffic will consist of video content (Cisco, 2019). The popularity of videos among private individuals is also evidenced by consumer interest in video services such as YouTube, which is the second most frequented website worldwide (Alexa, 2020). Additionally, video content has spurred many new businesses, such as TikTok, which leverage the popularity of moving images and have grown rapidly (Leskin, 2020). Corporate budget allocation decisions mirror the popularity of videos. For example, online video advertising budgets are projected to grow by 6% annually over the next four years (Statista, 2019).

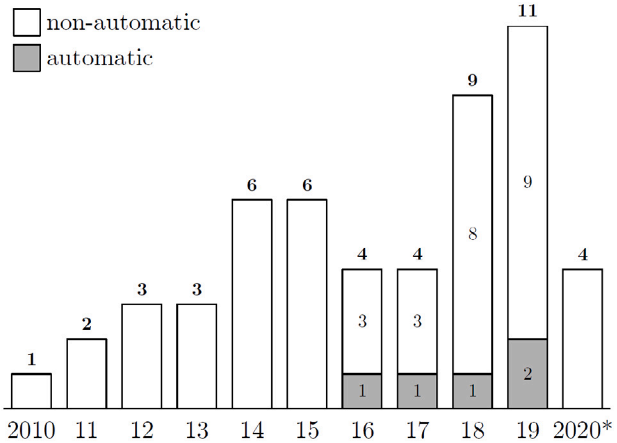
Accordingly, video formats have gained interest in business research in the past decade, leading to an almost continuous increase in the number of video-related papers published each year (see Fig. 1). Although manually analyzing videos can be a very cumbersome task and almost impossible to scale to larger datasets, only a few publications in business research on videos have employed automated analysis. Nonetheless, many prominent video-related applications have already emerged, e.g., predicting movie success (Eliashberg & Sawhney, 1994; Himes & Thompson, 2007), creating engaging video games (Wood et al., 2004), or explaining virality of commercials (Akpinar & Berger, 2017; Dessart & Pitardi, 2019; Simmonds et al., 2019; Tellis et al., 2019). Recently, a few researchers have started to work with automated video analyses (e.g., Li et al., 2019; Liu et al., 2018). Automated analysis is important because it allows studying more observations to investigate more nuanced effects, interactions, or rarely occurring (but potentially important) events. It also enables researchers to control for a larger variety of video features and avoid potential confounds or omitted variable bias. In addition, automated analysis facilitates research replications and extensions because it does not rely on subjective coding.

Videos consist of sequences of images, with typically 20 to 30 images (frames) per second (fps). The lower boundary of 20 fps roughly represents the threshold at which humans perceive a sequence of frames as fluid motion (Berkeley Institute of Design, 2012). A central component to any automated video analysis is therefore the analysis of individual images. In the past years, significant progress in the field of image mining has been made (e.g., Burnap et al., 2019; Schikowsky et al.,

☆ **Funding:** This work was funded by the German Research Foundation (DFG) research unit 1452, "Marketing of Products in the Age of Digital Social Media", HE 6703/1-2.

* Corresponding author.
E-mail addresses: jasper.schwenzow@uni-hamburg.de (J. Schwenzow), jochen.hartmann@uni-hamburg.de (J. Hartmann), amos.schikowsky@uni-hamburg.de (A. Schikowsky), mark.heitmann@uni-hamburg.de (M. Heitmann).

https://doi.org/10.1016/j.jbusres.2020.09.059
Received 23 January 2020; Received in revised form 25 September 2020; Accepted 27 September 2020
Available online 14 October 2020



Note.

The data base consists of all papers published in major business journals from 2010-2020 found by a systematic video-related keyword search in the abstracts

*Year not completed

Fig. 1. Number of Video Papers in Business Research.

2020), and some of these developments have been translated to moving images. However, we are not aware of a consolidated open-source video mining toolbox for business research. While some recent publications employ automated analyses (e.g., Li et al., 2019; Liu et al., 2018), these are limited to specific subsets of video-based features (typically from 2–3) and often do not provide source code or utilize proprietary software services. We therefore consolidate feature extraction theory from both image- and video-related research to build an open-source tool to extract and aggregate video features that are relevant for the types of econometric models investigated in business research.

Mining videos is a nontrivial task with many challenges, starting with data preparation and extending to final extraction. First, a selection of features for extraction must be made. This is a particular issue when manually coding videos, which is likely to be one of the reasons that research has proceeded with only a few features at a time so far. However, such a focus on selected features can create omitted variable bias and erroneous substantive conclusions since video consumption is a holistic experience. It is therefore desirable to find ways to automatically extract as many features as possible and make use of appropriate econometric techniques or machine learning to create dense representations and test their impact. This goal is complicated by the fact that information on how to obtain video features is widely scattered across diverse literature, which drastically increases implementation costs for business researchers. While several commercial services exist, these are mostly a *black box* to researchers and typically do not report classification accuracy. These services might also be subject to opaque changes and discontinuation by the commercial vendors, endangering key constituents of business research, namely, transparency and reproducibility.

In terms of technical implementation, the information richness of videos can make data processing cumbersome due to large file sizes and datasets. Hence, efficient approaches to (a) extract relevant features and to (b) aggregate them meaningfully from frame level to video level are needed so that data can be appropriately handled in econometric models.

This research structures and discusses interpretable video features, which can be currently extracted with open-source techniques, and consolidates them into a comprehensive analysis framework, including (a) extensions of image mining techniques to moving images, e.g., for face, emotion, and object detection, (b) implementation of formerly manually coded established visual concepts such as colorfulness, and (c) extensions and enhanced combinations of newly developed state-of-the-art techniques based on a systematic screening of video-related GitHub repositories, such as an embedding-based visual variation measure. We additionally implement functions to aggregate those features in a meaningful way. To facilitate application, we provide the consolidated Python scripts and concept implementations as one easy-to-use tool[1] with a Colab notebook acting as a graphical user interface (GUI), thereby requiring very little to no programming knowledge while enabling full transparency of the code. We also provide a step-by-step video tutorial to assist readers in terms of application.[2] Based on an exemplary case study on movie advertising, we illustrate how the extracted features can be included in econometric models to gain substantive insights. Specifically, we apply our tool to 975 movie trailers. For these data, adding static frame-based video features such as the presence of human faces increases the explanatory power by nearly 17% compared to a baseline model based on movie budget, release timing, and genre. Additionally, adding dynamic features based on frame sequences (e.g., scene cuts) improves the explanatory power by nearly 30%. These findings suggest that relevant effects can be detected with automated analysis. Similar procedures can be applied to a variety of business research problems to complement the growing body of image mining research (e.g., Hartmann et al., 2020; Li et al., 2019).

2. Video mining

2.1. Related literature

As shown in Fig. 1, there has been a significant increase in the number of papers analyzing videos in the past ten years. We identified the relevant literature by first selecting high ranking English language business journals[3], which are relevant to the field of video analysis. We then gathered a database of all papers published in these journals from May 2010 to June 2020 based on a Scopus extract, leading to a dataset of 10,742 papers. We then performed a keyword search on the abstracts of those papers, producing a subset of 215 papers.[4] We then proceeded manually, keeping all papers that specifically handle video-related features and excluding false positives (e.g., economic trends in the video game industry) by accessing and checking all remaining papers. This led to our final collection of 53 relevant video papers. The complete list of video papers is available in Web Appendix A.

The majority of these papers use video feature extraction to explain ad effectiveness (e.g., Couwenberg et al., 2017; Dessart, 2018; Guitart et al., 2018; Jeon et al., 2019; Tucker, 2015) or social media content virality (e.g., Akpinar & Berger, 2017; Liu-Thompkins & Rogerson, 2012; Shehu et al., 2016; Tellis et al., 2019). Other popular use cases include various analyses of human behavior, such as frontline staff interactions (e.g., Marinova et al., 2018), in-store consumer behavior (e.g., Hui et al., 2013; Zhang et al., 2014), or management performance (e.g., Choudhury et al., 2019; Gylfe et al., 2016). Moreover, previous research has employed video analyses to optimize video products such as movie clips (Liu et al., 2018) or TV shows (Hui et al., 2014).

The video features extracted across these papers can be divided into either structural or content features (see Table 1, Lang et al. (1993)). Structural features encompass all lower-level features, e.g., colors, scene cuts, and duration (Vijayakumar & Nedunchezhian, 2012), and can be typically altered while editing the footage to produce the final video and eventually determine *how* the video is presented. These are

[1] The tool is available on GitHub: https://github.com/JasperLS/Understanding_Videos_at_Scale/blob/master/Understanding_Videos_at_Scale.ipynb.

[2] http://youtu.be/DnAEHjdg6u8.

[3] We chose Journal of Marketing, Marketing Science, Management Science, Journal of Marketing Research, International Journal of Research in Marketing, Journal of the Academy of Marketing Science, Journal of Business Research, Journal of Interactive Marketing, and Strategic Management Journal.

[4] Keywords: video, visual media, tv, television, trailer, clip, spots, videos, commercials, tvs, clips.

Table 1
Extracted features in video literature.

	Automatic	Non-automatic
Content features	• Objects (e.g., Li et al., 2019) • Emotions (e.g., Choudhury et al., 2019; Lu et al., 2016)	• Humans, animals (e.g., Bellman et al., 2012; Dessart, 2018) • Objects (e.g., Kumar & Tan, 2015) • Emotions (e.g., Bellman et al., 2012) • Text content (e.g., Fossen & Schweidel, 2019a; Roberts et al., 2015) • Story (e.g., Akpinar & Berger, 2017; Loewenstein et al., 2011) • Branding, sponsorship (e.g., Tellis et al., 2019) • Message tone, e.g., functional, emotional (e.g., Geuens et al., 2011) • Other specific characteristics, e.g., stereotypes (e.g., Avraham, 2018)
Structural features	• Visual variation (e.g., Couwenberg et al., 2017) • Scene cuts (e.g., Liu et al., 2018) • Color characteristics (e.g., Couwenberg et al., 2017) • Duration (e.g., Li et al., 2019)	• Quality (e.g., Hautz et al., 2014; Liu-Thompkins & Rogerson, 2012) • Interactive elements, e.g., skip button (e.g., Jeon et al., 2019)

usually objective technical features and are often less substantively interesting but are important to control. In contrast, higher-level content features include *what* is presented, encompassing not only visual cues (faces, objects), but also automated predictions of viewer perceptions (emotions, story).

Interestingly, even papers with very similar or identical dependent variables have explored very different video features. For example, both Couwenberg et al. (2017) and Chandrasekaran et al. (2017) investigate ad effectiveness. However, Couwenberg et al. (2017) mainly analyze the effect of lower-level features such as luminance, visual variance, and scene cuts, while Chandrasekaran et al. (2017) focus on the higher-level functional vs. emotional tone of the ads, suggesting that more comprehensive statistical models can be attained with more comprehensive video coding.

With only five out of 53 video studies, less than 10% took an automated approach to extract features from videos (see Fig. 1). The number of extracted visual features per paper varied between one (Choudhury et al., 2019) and three (Li et al., 2019), representing a small subset of the large range of possible features. Regarding the types of features, both content features such as the presence of specific objects (Li et al., 2019) or emotions (e.g., Choudhury et al., 2019; Lu et al., 2016) and structural features, such as visual variation (e.g., Couwenberg et al., 2017; Li et al., 2019) and scene cuts (Liu et al., 2018) were successfully extracted.

Among the studies that employ automatic techniques, the majority (60%) work with proprietary solutions. For example, Choudhury et al. (2019) use the proprietary Microsoft API to extract facial expressions. While this API is based on a published algorithm by Yu and Zhang (2015), the parameter values and potential refinements are undisclosed and subject to modifications. Similarly, as there is no holistic feature extraction script available, Liu et al. (2018) use two different types of proprietary software to extract facial emotions and scene cuts. Li et al. (2019) go further in taking a mixed approach by measuring technical features but complete the image recognition work in collaboration with a third-party provider with proprietary models. While the results are useful, greater transparency would facilitate replication and extension. For example, if these commercial services cease to exist or are being adapted over time, other researchers would not be able to replicate the published findings.

In contrast, Couwenberg et al. (2017) used MATLAB to extract low-level features, such as luminance, and Lu et al. (2016) used it to detect faces. This is commendable in terms of replications, which could be further facilitated by providing the actual code utilized for feature extraction. While the vast majority of open-source video and image analysis tools are published in Python (e.g., at the time this paper was written, a search for “video analysis” on GitHub delivered more than 500 repositories using Python and 114 using MATLAB), no tools among our subset of papers utilized this popular programming language. Given the contributions that have already been achieved, there appears to be much potential in adding the publicly available feature extraction scripts that have not been tapped into by business research. The lack of an overview and the costs involved to evaluate and integrate the individual Python scripts have likely contributed to the low penetration of these open software approaches into business research.

2.2. Automatic feature extraction

Importantly, automatic feature extraction can be conducted on two levels. Most features, e.g., colorfulness, are extracted on the frame level and then aggregated to the video level, e.g., average colorfulness (see Fig. 2). Other features are only meaningful on the video level, such as scene cuts, visual variance, and video duration. In this regard, researchers dealing with video data face two challenges. They first need to select relevant features on the frame level. Then they need to find ways to meaningfully aggregate these features to the video level. On the frame level, features can be extracted independently for every frame of a given video; e.g., each frame has a certain colorfulness and can potentially have one or more faces on it. In terms of extraction methods, structural features typically do not require machine learning-based methods for extraction and are condensed in a single average value per frame, e.g., one average colorfulness score for each frame (Li et al., 2019).

On the other hand, content features typically require classification or content detection with machine learning to code videos at scale (e.g., employing deep convolutional neural networks to detect faces in moving images). While a classification provides a single value per frame, usually accompanied by a confidence level, a detection results in bounding boxes for each detected object (see the left part of Fig. 2). Consequently, there exist multiple ways to construct detection-based features, e.g., the number of faces, the size of faces, or the position of faces (Filntisis et al., 2019).

This challenge of parameter richness on frame level increases in significance when aggregating to the video level (Fig. 2). Some of the objective video characteristics relate directly to the video level, including metadata such as video duration or resolution. However, many video features are aggregations from frame-level measures. In many econometric models, the unit of analysis is the video. Hence, frame-level information – such as the presence of protagonists or distribution of colors – must be meaningfully aggregated to video level. Which method is appropriate depends on the substantive research question. Choudhury et al. (2019), for example, use the overall average of frame-level emotion scores to investigate how overall emotions relate to future merger and acquisition decisions. Liu et al. (2018) construct three parameters (total number of scenes, average length, location of the longest scene) to detect scene cuts and gain more granular insight into the ideal way to cut trailers. Such analysis on the individual scene or frame level allows studying the impact of different dynamics of video content. We found only a few publications that consider the dynamic

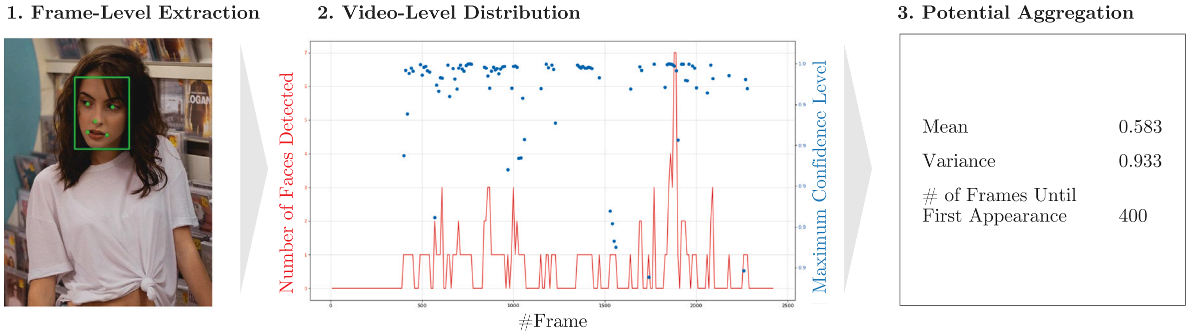


Fig. 2. Meaningful Aggregation from Frame to Video Level.

changes and relative positions of video features. This includes Choi et al. (2018) and van Reijmersdal et al. (2020), who consider the timing of certain events within a video, and Dessart and Pitardi (2019) and Hui et al. (2014), who analyze different sequences of events.

3. Video feature extraction

3.1. Tool creation

A growing number of researchers in business fields rely on increasingly complex programming to study unstructured data (e.g., Choudhury et al., 2019; Hartmann et al., 2019), making reproducible and transparent code essential so that further research can build upon these findings (Somers, 2018). To facilitate such research, we consolidate feature extraction tools and algorithms from multiple sources and provide a single, comprehensive Python tool for extracting 11 features resulting in 109 different variables. A complete variable description is available in Web Appendix B.

We took three steps to create this single video mining tool.[5] The objective of creating this tool is to facilitate extraction of a comprehensive set of video features. We want to enable researchers to fully leverage the technological potential for substantive insights while ensuring that the results can be replicated. To achieve these objectives, we first searched GitHub for everything labeled as 'video analysis'. We excluded all results not available in Python since we wanted to build a coherent tool based on the language that by far had the most results due to its prominent role in machine learning (KDnuggets, 2017). Next, we manually checked whether the resulting list of tools would cover the identified relevant variables for business research purposes. If multiple tools for the same feature were available, we implemented them based on accuracy on our test data, which also correlated with ratings on GitHub. Next, we systematically searched the web outside GitHub for tools that would enable extraction of the features still missing after the previous steps, including the detection and recognition of humans, animals, various objects, and text content. Finally, we augmented our tool by adding features that were not covered up to this point but were mentioned in related business research and technically possible to extract via the tool, including visual variation, quality, color features, duration, and resolution.

3.2. Implementation

The tool is built in Python and consists of three parts: (1) an easy-to-use notebook in Colab, which acts as the GUI and both collects user input and executes all lower-level scripts, (2) a feature extraction script called 'feature_extraction_main.py', which loops over all videos and extracts the features, and (3) all required materials, including additional modules, model weights, and other supplementary data. To extract video features, our tool only requires a researcher to have an internet connection to open the Colab link and click through our tool following the instructions. Accordingly, our GUI approach enables users without programming knowledge or without the specific hardware to extract video features in an accessible way without the cumbersome task of setting up a dedicated programming environment. However, all scripts are open-source Python code and can be transparently accessed and adapted to individual user needs. The tool is also available for download to run on local machines.

Fig. A.1 presents the processing pipeline of the feature extraction tool. The input is collected in the Colab notebook, which then calls the main script for feature extraction. The script automatically codes the videos located in a given folder from a start to an end index and saves all extracted feature values to .csv files in a provided Google Drive output folder that can be used for further econometric modeling.

All information is also available in our annotated GitHub repository. We additionally provide a simple YouTube tutorial on how to use our code.

3.3. Processing time

When working with large numbers of videos or very long videos, the processing time can become a limiting factor. This can be addressed with multithreading and parallelization. As a simple first solution, we advise researchers to split their videos across multiple folders and simply start multiple Colab notebooks in parallel. Furthermore, reducing analytical detail improves efficiency. For example, researchers focus on parts of the images, e.g., color characteristics only on a random sample of the pixels or subsampling to every nth frame. As there are typically more than 20 frames per second (fps), one can analyze every 5th frame with only minor loss of information.

Choudhury et al. (2019) even sample only one image frame per second.[6] The empirical application that we report below is based on 975 videos (average length of 2:33 min) that we coded in 48 hours while applying 5 batches in parallel on Colab, using two GPU and three CPU notebooks. This demonstrates that even large video collections can be classified with reasonable computation times.

[5] The tool is available on GitHub: https://github.com/JasperLS/Understanding_Videos_at_Scale/blob/master/Understanding_Videos_at_Scale.ipynb.

[6] If significantly more detail reduction is required, we found it useful to combine to leverage scene cut information, e.g., extracting color information from only three frames per scene.

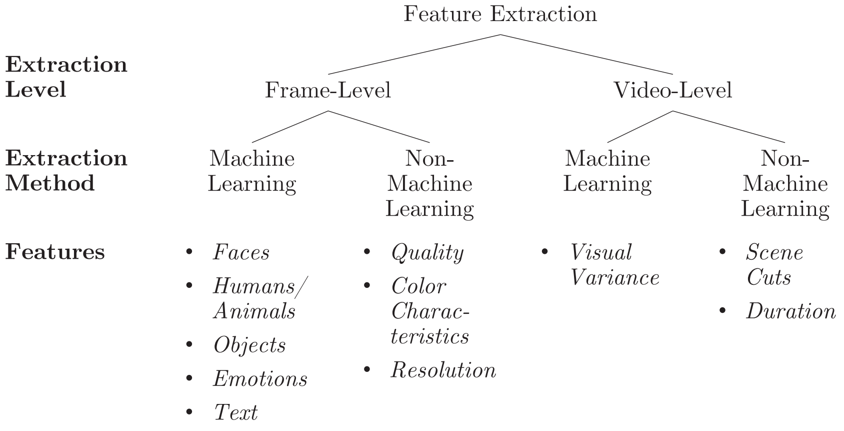


Fig. 3. Features & Extraction Methods.

3.4. Video features

Fig. 3 summarizes all video features from the group of potentially relevant features identified in Table 1 that we were able to find as well-performing and open implementations. From these features, we can extract variables (i.e., mathematical objects for modeling) such as scene cut frequency, representing the average number of scene cuts per second.

Following our previous description, features are first either extracted at the frame or video level. Second, extraction methods can be distinguished in machine learning and non-machine learning. Since frame-level dependent variables are seldom available, all frame-level features typically require some form of video aggregation to perform econometric analysis of video performance. Our tool provides both frame-level results for those features and potentially relevant basic aggregations, allowing for quick results while providing the option of choosing different aggregation methods if required. See Table B.1 for a detailed description of the output.

3.4.1. Classification and detection performance

Video features, which require classification or detection methods, especially machine learning, are more difficult to obtain. These methods are often trained to predict manually annotated data, which produces classification errors. This in turn is a function of the type and quality of the video material of interest. Therefore, performance measures should be reported due to their potential impact on substantive interpretations (e.g., Hartmann et al., 2019). To provide confidence in the results, we compared the results of our tool for each classification or detection feature against a human coder on a random (hold-out) sample of 20 movie trailers. Following conventions to measure machine learning performance in applied research (e.g., Netzer et al., 2012), we use precision and recall as our primary performance measures. Precision describes the proportion of correctly classified entities over all classified entities, and recall describes the proportion of correctly classified entities over all potentially correct entities, e.g., for scene cuts:

$$precision = \frac{|\{detected\ scene\ cuts\} \cap \{actual\ scene\ cuts\}|}{|\{detected\ scene\ cuts\}|} \quad (1)$$

$$recall = \frac{|\{detected\ scene\ cuts\} \cap \{actual\ scene\ cuts\}|}{|\{actual\ scene\ cuts\}|} \quad (2)$$

The results are shown in Table 2. For scene cuts, faces, and humans, our tool achieves relatively high scores between .75 and .95 for precision and recall and > .80 for the respective F1 scores. For the different emotions, both precision and recall fluctuate, which is not surprising, given the difficulty of consistently classifying emotions, even among human coders (Barrett et al., 2019).

Table 2
Precision & recall values.

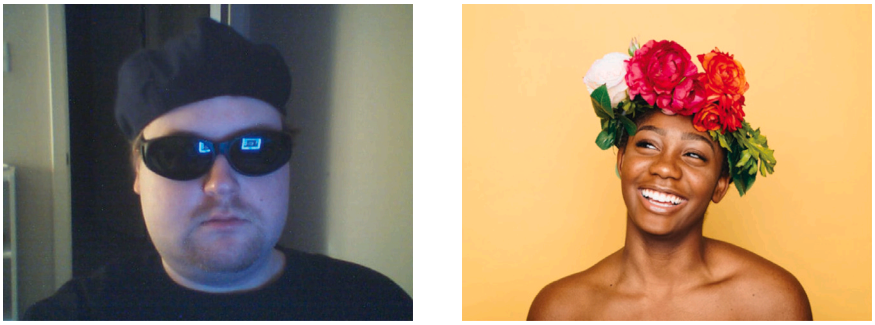
Feature	Precision	Recall	F1 Score
Scene Cuts[a]	0.983	0.821	0.895
Humans[b]	0.904	0.945	0.924
Faces[b]	0.914	0.757	0.828
Anger[c]	0.323	0.278	0.299
Disgust[c]	0.273	0.083	0.128
Fear[c]	0.250	0.333	0.286
Happiness[c]	0.633	0.861	0.729
Sadness[c]	0.571	0.111	0.186
Surprise[c]	0.567	0.583	0.575
Neutrality[c]	0.348	0.667	0.457

Note.
[a] Tested on a random, manually coded sample of 20 trailer.
[b] Tested on a random, manually coded sample of 200 images from 20 trailer.
[c] Tested on a public dataset of labeled faces.

3.4.2. Frame-level features

Faces. Since human faces attract attention (Tomalski et al., 2009), faces are often visible in traditional advertising (Xiao & Ding, 2014). Similarly, social media research suggests, faces result in stronger engagement and emotional response (Bakhshi et al., 2014). Detection of faces is also critical because face detection is the basis for other more advanced analysis (e.g., on human presence in general, facial emotions). To ensure transparency and replicability, we suggest using open-source software, such as MTCNN face detection (Zhang et al., 2016). Depending on the desired research objective, there are many ways to aggregate this frame-level analysis. Obvious starting points include the *share of frames with faces* (i.e., frames with at least one face out of all frames) and *average number of faces* (i.e., faces per frame averages across all frames). Many further aggregations are conceivable, e.g., when the first face appeared, sizes and positions of faces, their amounts, and types of movement, which can be computed based on the output of the tool presented in this paper.

Other Objects. Similar to face detection, popular open-source solutions exist for object detection, e.g., the YOLOv3 architecture with 80 of the most common objects, including human bodies, animals, and various everyday items (Lin et al., 2014). The YOLOv3 architecture is pretrained on COCO, which is the most popular dataset for object detection (Iqbal et al., 2018). The YOLOv3 neural network with 53 layers produces bounding boxes around each object making both exact locations and object sizes available (Fukushima, 1975; LeCun et al., 1989). Aside from the type of objects, their number and frequency of



(a) Low-quality portrait
Quality measure: 47.23

(b) High-quality portrait
Quality measure: 179.43

Fig. 4. Comparison of Frame-Level Quality.

occurrences can be computed based on this information as well as the types of variables also relevant to face analysis.

Emotions. Numerous studies have demonstrated the relevance of emotions to the effects of advertising (e.g., Teixeira et al., 2010; Tellis et al., 2019) or used them to explain managerial performance (Choudhury et al., 2019). To retrieve facial emotions, we used a publicly available implementation from GitHub, which uses a relatively small artificial neural network consisting of five convolutional and three fully connected regular layers. We test the performance on a publicly available dataset of labeled faces.[7] According to the results, happiness and surprise can be reliably detected, while the detection of other emotions should be treated with caution (Table 2). Since performance can also vary across faces and context, it is important to assess and report precision and recall for every application to avoid misleading substantive interpretations of the effects of individual emotions.

Quality. The quality of images is an important part of the visual presentation and can lead to major differences in demand (Zhang et al., 2017). As a measure for quality, we evaluate the focus in the frames. There is a large body of literature on how to find the best focus, and their methods can be used to evaluate it. We used the implementation from Pech-Pacheco et al. (2000) due to its simplicity, speed, and empirical performance, as seen in Fig. 4.

The idea behind this method is the assumption that high-quality images have regions with sharp edges as well as regions without large differences in intensity. The Laplace filter – a discretization of the Laplace operator (sum of the second derivatives) – is used to detect edges as a convolution[8] with the following kernel:

$$\frac{1}{6}\begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

A high variance of the obtained result indicates both the presence of sharp edges and areas without edges, which typically coincides with proper lighting and a high-quality camera (no blurriness or optical aberration). Accordingly, we use it as an indicator of quality images.

Color. Research has emphasized the role of color in human perception and response (Schindler, 1986). Colors can be represented in the RGB color space (an additive mixture of the colors red, green, and blue). However, aggregation on the frame level is nontrivial since three equivalently sized patches of pure red, green, and blue would produce the same score as a gray frame. An alternative color space is the hue, saturation, and value (HSV) color space. Hue is the degree to which a color stimulus is similar to red, green, blue, and yellow (Elliot et al., 2015). However, when taking averages across an image to determine color utilization, the result can no longer be interpreted as the colors are mixed. To extract the total display of specific colors per frame, we use a complementary measure, which is trained to infer color from fuzzy labeled real-world images to construct statements of color utilization (Joost et al., 2009). Our tool uses this measure to extract the average display of the colors black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow, as they are perceived by humans. Potential further colors can be added.

Colorfulness. While *Color* captures the perceived presence of specific colors such as red overall colorfulness is captured in a colorfulness variable. Even with an equal display of a specific color, light conditions, and other factors of a frame can still result in a higher colorfulness. We implemented colorfulness as described in Hasler and Süsstrunk (2003).

Saturation. Saturation refers to the degree of difference from a *pure* color to the respective gray having the same luminosity.

Value. Value is the luminosity — where 0% refers to pure black and 100% is the brightest color with a given hue and saturation (*H* and *S* channels, respectively). Compared to brightness, which can be calculated as the average of the three channels R, G, and B, it more closely resembles human brightness perception.

Resolution. In terms of non-machine learning frame-level measures, the resolution of the video can easily be derived from the matrix shape of each frame and represents the number of pixels per height and width per frame. When computing such measures, black stripes for reduced frame sizes need to be addressed.

3.4.3. Video-level features

Visual variance. Similar to scene cuts and other structural complexity drivers, it can be assumed that visual variance is fundamentally important because it leads to arousal, which may affect dependent variables such as attention and memory (Lang et al., 2007). Li et al. (2019) have shown that increasing visual variance can have a positive effect on the success of crowdfunding campaigns. As we did not find a publicly available script on visual variance, we implemented a deep learning-based solution ourselves, combining existing publicly available tools. Specifically, we use the middle frame from each scene within the video to calculate two measures for the visual variance of the video, namely,

[7] The dataset is available at http://app.visgraf.impa.br/database/faces/ and includes 7 facial emotions of 38 different people.

[8] See the Appendix of Schikowsky et al. (2020) for an introduction into how convolutions work.

(1) the variance from scene to scene and (2) an average of this value across all scenes, to obtain video-level visual variance.[9]

To calculate the visual variance between two images, we used a Siamese network approach (Koch, 2015) using ResNet-152 (He et al., 2016), a deep artificial convolutional neural network that is pretrained on the task of image classification on the ImageNet dataset (Deng et al., 2009). For each pair of images, we applied it to both images separately and extracted the last layer before the classification layer. This last layer contains the high-level features as a 2,048-dimensional vector, including information on colors, shapes, objects, and textures, and their sizes and locations. We then calculated the angular distance (cosine similarity) in the embedding space as a measure for visual variance. The advantage of this approach is its speed of execution, as we only need to calculate the embedding of a single frame per scene while still retaining most of the information of the whole scene. At the same time, the artificial neural network captures differences in colors, shapes, objects, and textures and their sizes and locations across frames.

Scene cuts. The most prominent video-level feature is the number of scene cuts. These have various applications, including detection of certain events that are typically accompanied by changes in scenes (e.g., Almousa et al., 2018; Assfalg et al., 2002). Initial evidence suggests that the types of scene cuts in a video can have important effects on video success (Liu et al., 2018). Scene cut detection can be accomplished with the open-source solution PySceneDetect (Almousa et al., 2018; Shou et al., 2018), which is integrated into our tool. The algorithm fundamentally compares the hue, saturation, and value of two neighboring frames and compares the average distances against a threshold. We tested this method on a random set of trailers and achieved satisfying results, with a precision of 0.98 and a recall of 0.82 (see Table 2).

It should be noted that a scene cut refers to all cuts, meaning that one scene can have multiple cuts, e.g., a filmed dialogue with iterations in perspectives. Various variables can be constructed based on this measure. Our tool includes *average scene cut frequency* as an intuitive and relevant aggregation (i.e., number of scene cuts per video second). Based on the output of the tool to this article, other variables can be constructed, such as the first scene cut, the distribution of scene cuts, acceleration, or deceleration of scene cuts.

Duration. Measuring video duration is straightforward but is nevertheless an important driver in various applications that should be controlled for in any analysis (e.g., Goodrich et al., 2015; Li et al., 2019; Quesenberry & Coolsen, 2019; Tellis et al., 2019).

3.5. Further options

We implemented two further options in our tool, which we considered helpful for business research: (1) We added the possibility to retrieve and apply feature extraction to online videos; and (2) We provided the additional possibility to extract various forms of text content. (1) We use the publicly available youtube-dl,[10] which – despite its name – allows downloading videos from not only YouTube but also a broad range of more than 1,000 popular websites with video content, including Facebook and Instagram. (2) Established optical character recognition (OCR) tools such as the open-source tesseract Smith (2007) are trained to perform well on nondistorted black text on a white background and hence must not perform well in the context required for the various forms of text that might occur in real-life video content, also referred to as scene text recognition (STR). However, two high-performing STR solutions have recently become available: one to localize characters (Baek, Lee et al., 2019) and one to recognize them (Baek, Kim et al., 2019), and both also function for in-the-wild text in images. We coupled both approaches and implemented this advanced STR module in our framework. We further augmented these functionalities by adding additional logic to rebuild sentences from detected characters, e.g., allowing for capturing multicolor promotional messages in a video clip or movie subtitles.

4. Empirical application: Online movie trailers

To illustrate how video mining can be applied in business research, we analyze a set of 975 online movie trailers using our tool. This empirical application relates to business research regarding movie success (e.g., Clement et al., 2014; Eliashberg et al., 2007; Eliashberg & Sawhney, 1994), which studies various (well-structured) explanatory variables such as budgets, marketing mix, or star ratings of actors. Extracting structured data from unstructured video content may provide incremental explanatory power. For illustrative purposes, we focus on movie trailers as opposed to full movies because these are shorter, easier to handle, and more homogeneous. At the same time, movie trailers are arguably one of the most important marketing tools when releasing a blockbuster movie (Bhave et al., 2015). We test whether video mining creates incremental value, i.e., whether extracted features impact performance over and above traditional (structured) covariates such as *budget*, *runtime*, and *movie age*. One of the main motivations to distribute online movie trailers is to generate interest in the movie. We approximate this by predicting the number of likes each movie trailer generates online as our primary dependent variable. For illustrative purposes, we apply the supplementary tool of this article without further feature engineering in a straightforward manner.

Data description. The initial dataset consists of the top 1,250 movies of the years 2016, 2017, and 2018 from the Internet Movie Database (IMDb), ranked by total US Box office success. We then filtered for movies with the original language English and for the availability of official trailers, which reduced our dataset of trailers to 975. For these trailers, we collect the total number of *likes* as well as the aforementioned other movie-related information based on IMDb (see Table C.1 for data descriptions and summary statistics). On average, each video attracts 40,972 likes. In terms of *genre*, 55.0% are *drama*, 18.1% are *comedy*, 11.9% are *thriller*, 5.4% are *action*, 0.6% are *romance*, and 9.0% are *other*. The average *IMDb rating* of all movies is 5.91 (SD: 1.81).

Modeling. In total, the supplementary tool to this article extracts 109 video variables (see Table C.1). The extracted variables have potential nonlinear and interactive effects on observer responses. For illustrative purposes, we focus on potential simple nonlinear (quadratic) effects for all (mean-centered) variables. Since our analysis is explorative, we employ parameter shrinkage (Lasso) to identify relevant predictors of movie trailer success. Doing so, we log-transform the dependent variable to account for the skewed variable distribution (see also Li & Xie, 2019) and standardize all feature variables, as Lasso regressions place a penalty on the magnitude of the coefficients associated with each variable, and is therefore affected by variable scales.[11] All VIF values are below 2, suggesting that multicollinearity is not an issue.

[9] We could also measure visual variance across all frames from a video; however, within a scene, the visual variance typically does not change significantly.

[10] Available at https://youtube-dl.org/.

[11] Instead of including all colors, which our tool extracts at the per-frame level, we include colorfulness as an aggregate measure in our regression. Moreover, of the 80 object categories (humans, animals, etc.) that we detect per frame, we include only humans because the presence of actors in a trailer is likely to be the most important object category. We also excluded quality, as all movie trailers were professionally produced and yielded few differences in quality.

Table 3
Regression results: Movie trailers.

	Dependent variable: Log-Likes		
	Model (1)	Model (2)	Model (3)
Content			
Video-Level			
Average Scene Cut Frequency			.42***
Average Scene Similarity			.05
Length			.48***
Length (sq.)			−.05
Frame-Level			
Human Area Coverage		−.44***	−.42***
Human Area Coverage (sq.)		−.28***	−.11**
Share of Frames with Faces		.03	−.11*
Average Saturation		.02	−.12**
Average Value		−.13	−.18**
Average Value (sq.)		.09	.25***
Average Colorfulness		−.22***	−.24***
Average Colorfulness (sq.)		−.10*	−.09*
Anger		−.10**	−.07
Fear		.19***	.18***
Surprise		.07	.08
Controls			
Budget	.41***	.32***	.31***
Runtime	.29***	.21***	.13**
Movie Age	.03	.07	.07
IMDb Rating	.58***	.51***	.40***
Views	.91***	.86***	.83***
Genre, Baseline: Action			
Comedy	.70**	.54*	.59**
Drama	.36	.21	.43
Romance	−.13	−.20	.03
Thriller	.81**	.62**	.56**
Other	−.40	−.35	−.07
Constant	8.35***	8.48***	8.33***
Observations	975	975	975
R^2	.48	.56	.62
Akaike Inf. Crit.	3,833.27	3,690.58	3,550.07

Note. $^{*}p < .10$; $^{**}p < .05$; $^{***}p < .01$.
All numeric variables are mean-centered.

Results. We first examine whether video mining can add explanatory power in econometric models of business research by comparing a model with only conventional covariates (Model 1) with one model including the aforementioned aggregated frame-level variables (Model 2) and one model with both aggregated frame- and video-level variables (Model 3). Results are shown in Table 3. All models control for the attainable likes and overall interest by controlling for the total number of video views (Swani & Milne, 2017). Additionally, all other conventional control variables have effects in the expected direction. Specifically, a larger *budget* (.41, $p < .01$), longer *runtime* in cinemas (.29, $p < .01$), more *views* (.91, $p < .01$), and higher *IMDb rating* (.58, $p < .01$) are positively associated with the number of *likes*. In terms of *genre*, *comedy* and *thriller* movies on average receive more *likes* than *action* movies (.70, $p < .05$ and .81, $p < .05$, respectively). With video mining, we were able to extract information that is relevant over and above these factors. The increase in explanatory power suggests that including video mining features benefits modeling; i.e., R^2 improves by nearly 30% (from .48 in Model (1) to .62 in Model (3)) when including all features. When we limit the analysis to only frame-level features (Model 2), the model fit (R^2) still improves by approximately 17%. Note that these estimates are conservative since we did not take two-way or higher level interactions into account and did not apply any additional feature engineering. These results suggest that even simple econometric models can benefit from the information contained in the extracted video features.

Video-level features. Regarding video-related features, the Lasso model identifies two critical video-level variables (Model (3)). First, *scene cut frequency* plays a role such that trailers with more frequent scene changes are associated with more likes (.42, $p < .01$). This correlation indicates that, for example, an increase by 15 scenes is associated with 5% more likes. We caution that scene cuts may be correlated with other unobserved variables so we cannot draw causal conclusions from this analysis. However, even minor video content changes have been found to have tangible effects. For example, Lang et al. (1993) find that more cuts increase short-term attention. Lang et al. (2007) further suggest that higher structural complexity of media may lead to increased liking. Similarly, Liu et al. (2018) demonstrate how changing scene selection and cutting can impact consumer reactions to movie clips and they quantify the business impact for Netflix trailers. Second, consistent with Berger and Milkman (2012), video duration has a positive effect, possibly because longer videos contain more information and therefore offer greater overall potential for arousal.

Frame-level features. The Lasso model further identified multiple relevant frame-level variables. Interestingly, our results suggest a negative association between the presence of people and human faces in movie trailers and liking. These results are consistent for the *human area coverage* across all frames (−.06, $p < .01$) as well as the *share of frames with faces* (−.01, $p < .1$). Note that these findings must be interpreted based on the distribution of the actual data. Based on this illustrative analysis, it is not clear whether these findings generalize to other movie trailers.

Beyond substantive narrative elements, objective technical features have several nonlinear effects. Specifically, *average saturation* is negatively linked to *likes* (−.12, $p < .05$). This may be due to highly saturated videos being too bright to attract likes on similar levels as lower-saturated videos. The *average value*, i.e., the luminosity, exhibits a U-shaped relationship (linear effect of −.18, $p < .05$ squared effect of .25, $p < .01$) such that low and high values can be typically found in videos with more likes. A potential reason for this effect might be that brightness is linked to emotion, with a high brightness encouraging positive emotions and a low brightness inducing negative emotions (Lakens et al., 2013). Because a movie trailer aims to arouse the viewer, both extremes may eventually positively influence liking. The *average colorfulness*, on the other hand, is negatively linked to likes, with diminishing marginal returns (linear effect of −.24, $p < .01$, squared effect of −.09, $p < .01$). This might suggest a critical range of colorfulness values with a stronger impact on trailer preferences.

This empirical application is clearly explorative in nature, illustrating the additional potential of video mining to explore video-content features. We lack a more comprehensive set of covariates and did not control for endogeneity, which limits our ability for causal inferences. However, the analysis suggests that the incorporation of simple video features has the potential for interesting findings across different features types. Depending on research objectives, further content-related categories can be added. Additionally, different aggregations of the frame-level features are conceivable (e.g., time until first appearance of certain objects), which may further improve explanatory power. Given the findings and the various extension possibilities, it appears fair to conclude that automatic feature extraction can both improve predictive power and generate meaningful substantive insights that would not be attainable through manual coding at reasonable costs.

5. General discussion

The proliferation of video content across domains makes analyzing videos at scale increasingly relevant for business research. At the same time, increasing computing power and the availability of advanced machine learning make such analysis feasible. However, not all relevant methods for feature extraction are readily accessible to business research. These are rather scattered across various types of publications. We aimed to provide an overview of how to automatically construct several important video features that have had an impact in our illustrative application. The supplementary tool to this article provides easy access to fully transparent and replicable methods of video mining. It requires no programming knowledge and is immediately available to any researcher with an internet connection. This approach represents a novel synthesis of various feature extraction techniques, offers a systematic method for analyzing videos, and can be employed as a feature extraction mechanism across contexts, ensuring replicability, accessibility, and adaptability (in terms of custom feature aggregation). As the empirical illustration on movie trailers illustrates, even its most straightforward application has potential for relevant insights. We acknowledge that this is only correlational evidence, but these results do demonstrate the apparent potential of video mining for business research. Since such research is still in its infancy, even the first steps based on easily accessible video features are likely to uncover further knowledge in the respective fields.

Videos naturally provide an abundance of information; i.e., many variables can be extracted that will likely interact in complex ways. This makes parameter reduction an important objective. Various approaches are available to accomplish this aim. Often, only a smaller set of variables will be critical. The results of our empirical illustration suggest that a large portion of the variance of movie trailer preferences can be explained with few variables and simple nonlinear effects. As shown with the COCO dataset, it becomes increasingly easy to detect a multitude of objects. To make use of the extracted information, researchers need to develop and standardize approaches to deal with these video features. We expect this goal to be achieved in future research by a mixture of applying theory-driven parameter selection, machine learning, and econometric modeling, which can be based on the features extracted by the supplementary tool to this research.

It will be of increasing importance for researchers to pay attention to current developments in computer science, as new tools for extracting more sophisticated features are constantly being developed. From a business research perspective, it will be specifically interesting to employ developments for extracting interframe features, such as the detection of jokes, plot twists or dramatic climaxes, and overall story development. Given the modular nature of our supplementary tool, such further features may easily be added.

Another challenging problem for this field of research will be the growing use of APIs for feature extraction, as observed in recent publications. While APIs may provide easy access to state-of-the-art feature extraction methods, their results might not be reproducible over time. The severity of this nonreproducibility will likely depend on the types of features of interest. For objective technical features (e.g., image resolution), it may not matter much which feature extraction method is used. Whenever machine learning is involved (e.g., object detection, face detection, detection of emotions), different algorithms can produce different types of error that may be related to the substantive question of interest. For these features, transparent reports of accuracy levels, potential biases, and reproducible open-source applications are essential.

Our extensive literature review across ten years of business research has suggested various research opportunities. Overall, we see four major fields of potential video mining use cases in business research contexts:

1. Video advertisements (e.g., television, online banners, influencers)
2. Video products (e.g., video games, movies, documentaries, sports events)
3. User-generated content (e.g., on TikTok, Twitter, Instagram)
4. Analyses of human behavior (e.g., staff interactions, live customer reactions, managerial appearances, recruiting)

While several studies exist in the field of television advertising and video virality, many of the other formats have received less attention in business research. Additionally, many of the existing studies have focused on relatively small video datasets considering the availability of videos. This may be due to a lack of scalable video mining technologies. As these become easily accessible, insights from both new video formats and significantly larger datasets are likely to emerge. This is important because the number of variables (and potential interaction effects) is much larger using tools such as that proposed herein. Larger datasets allow detecting more nuanced effects and modeling less frequently occurring phenomena.

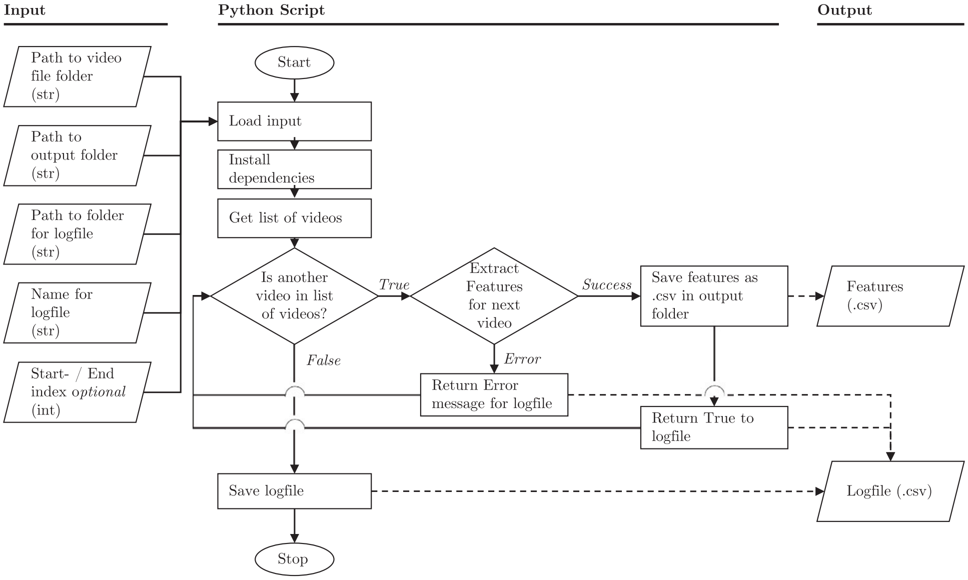
Beyond the computational challenges that arise, we find that the main two barriers that researchers must overcome are (a) extracting a relevant set of frame-level features and (b) meaningfully aggregating them on the video level for subsequent econometric analyses. The video mining tool provided with this research can simplify these steps and can be extended in several ways. As very few publications have examined the importance of different dynamics of video features within one video, researchers may use alternative aggregations of the variable output of our tool, e.g., rising or falling scene frequency, to explain consumer responses.

Furthermore, we have focused on visual content in this research. Various additional insights are possible when classifying audio content (Hildebrand et al., 2020). This includes interactions between the visual and audio content of videos, e.g., semantical synchronicity. Moreover, it may soon be possible to represent videos as low-dimensional video embeddings in a similar spirit as Word2Vec for text content (Mikolov et al., 2013). Such advances would unlock the possibility to analyze the similarity between different videos and understand which clusters exist. The visual variance measure that we propose is based on the scene-by-scene comparison of such deep learning embeddings. Overall, we hope that this research encourages other researchers to embrace the business research opportunities that arise with the advances in video mining technologies and the increasing proliferation of video content.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A


Input
Python Script
Output
Path to video file folder (str)
Path to output folder (str)
Path to folder for logfile (str)
Name for logfile (str)
Start- / End index optional (int)
Start
Load input
Install dependencies
Get list of videos
Is another video in list of videos?
True
Extract Features for next video
Success
Save features as .csv in output folder
Features (.csv)
False
Error
Return Error message for logfile
Return True to logfile
Save logfile
Logfile (.csv)
Stop

The script is accompanied by a step-by-step video-based tutorial that walks the viewers through each step from setup to output.

Fig. A.1. Flowchart of Python Tool.

Appendix B

Table B.1
Description of outputs from video mining tool.

Name	Features	Description
_FrameLevel_colors.csv	colorfulness, saturation, value, multiple colors	one line per analyzed frame, includes average values, scaled from 0 to 1, with 1 representing the maximum (e.g., 1 for red indicates a completely red screen for this specific frame)
_FrameLevel_embeddings.csv	–	one line per analyzed frame, includes 2048 dimensional embedding from last layer of ResNet-152 architecture, pre-trained on imagenet.
_FrameLevel_emotions.csv	emotions	one line per detected face, includes frame value and confidence scores for each emotion scaled from 0 to 1 (e.g., 0.85 for angry represents a face which was classified as angry with 85% confidence)
_FrameLevel_faces.csv	faces	one line per analyzed frame, includes confidence score scaled from 0 to 1 and pixel values of bounding box and facial features for each face detected
_FrameLevel_objects.csv	objects, humans	one line per detected objects, includes pixel values of bounding box, confidence score and type of object detected
_FrameLevel_quality.csv	quality	one line per analyzed frame, includes quality score from small numbers for low quality (blurry) to high numbers for high quality (no blurriness)
_FrameLevel_scenes.csv	scene cuts	one line per detected scene, includes frame and time code of start and end, and length in frames and seconds
_FrameLevel_similarities_neighbor.csv	visual variance	similarities between neighboring scenes (row n has cosine similarity between scene n and $n + 1$
_FrameLevel_similarities.csv	visual variance	symmetric matrix A of scene similarities between all scenes (matrix element $a_{i,j}$ has cosine similarity between scene i and j
_VideoLevel_colors.csv	colorfulness, saturation, value, multiple colors	one line per feature, includes mean and variance of each feature for all analyzed frames
_VideoLevel_emotions.csv	emotions	one line per emotion, includes mean and maximum of confidences of all detected faces for each emotion
_VideoLevel_faces_avg_number.csv	faces	includes the average number of detected faces of all analyzed frames, that included at least one face
_VideoLevel_faces_ratio.csv	faces	includes the share of frames with at least one detected face over all analyzed frames
_VideoLevel_length.csv	duration	includes the length of the video in seconds
_VideoLevel_objects_human_area.csv	humans	includes the average area share of the screen that is covered by people on all frames with people
_VideoLevel_quality.csv	quality	includes the average quality across all analyzed frames
_VideoLevel_resolution.csv	resolution	includes the resolution in pixels (height x width)
_VideoLevel_scenes_freq.csv	scene cuts	includes the average number of scenes per second
_VideoLevel_similarities_all.csv	visual variance	includes the averages and variances of neighboring and all scene similarities

Appendix C

Table C.1
Variable overview.

Variable	Description	Mean	SD
Video			
Video-Level			
Average Scene Cut Frequency	Total number of scenes divided by video length in seconds	.60	.15
Average Scene Similarity	Average cosine similarities between all scene embeddings	.77	.01
Length	Length of video in minutes	2.33	.42
Frame-Level (Aggregated)			
Human Area Coverage	Average share of pixels belonging to people in %	20.21	7.48
Share of Frames with Faces	Number of frames with faces divided by the number of frames in %	36.10	15.63
Average Saturation	Average saturation taken from S channel in HSV color space in %	32.80	9.16
Average Value	Average luminosity taken from V channel in HSV colorspace in %	24.26	8.88
Average Colorfulness	Average colorfulness as defined in Hasler and Süsstrunk (2003) in %	25.29	8.91
Anger	Average confidence level for angry across all faces in %	.12	.05
Fear	Average confidence level for fear across all faces in %	.16	.06
Surprise	Average confidence level for surprise across all faces in %	.03	.04
Controls			
Budget	Budget as indicated on IMDb in million US$	23.90	47.67
Runtime	Runtime in cinemas in months	8.69	1.42
Movie Age	Movie age in years	2.80	.91
IMDb Rating	Average user rating on IMDb (from 1 to 10)	5.91	1.81
Views	Number of views in million	6.65	12.79

Note. Resolution of video as pixel width times pixel height is constant across our sample at [720;1080].

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbusres.2020.09.059.

References

Akpinar, E., & Berger, J. (2017). Valuable virality. *Journal of Marketing Research*, *54*(2), 318–330. http://dx.doi.org/10.1509/jmr.13.0350.

Alexa (2020). The top 500 sites on the web. https://www.alexa.com/topsites.

Almousa, M., Benlamri, R., & Khoury, R. (2018). NLP-Enriched automatic video segmentation. In *2018 6th international conference on multimedia computing and systems* (pp. 1–6). IEEE Computer Society, http://dx.doi.org/10.1109/ICMCS.2018.8525880.

Assfalg, J., Bertini, M., Del Bimbo, A., Nunziati, W., & Pala, P. (2002). Soccer highlights detection and recognition using HMMs. In *Proceedings - 2002 IEEE international conference on multimedia and expo: Vol. 1* (pp. 825–828). Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.1109/ICME.2002.1035909.

Avraham, E. (2018). Nation branding and marketing strategies for combatting tourism crises and stereotypes toward destinations. *Journal of Business Research*, http://dx.doi.org/10.1016/j.jbusres.2018.02.036.

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S. J., & Lee, H. (2019). What is wrong with scene text recognition model comparisons? Dataset and model analysis. In *International conference on computer vision*.

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9365–9374).

Bakhshi, S., Shamma, D. A., & Gilbert, E. (2014). Faces engage us: Photos with faces attract more likes and comments on instagram. In *Conference on human factors in computing systems - proceedings* (pp. 965–974). Association for Computing Machinery, http://dx.doi.org/10.1145/2556288.2557403.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. http://dx.doi.org/10.1177/1529100619832930.

Bellman, S., Schweda, A., & Varan, D. (2012). Interactive TV advertising: ITV ad executional factors. *Journal of Business Research*, *65*(6), 831–839. http://dx.doi.org/10.1016/j.jbusres.2011.01.003.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205. http://dx.doi.org/10.1509/jmr.10.0353.

Berkeley Institute of Design (2012). Human information processing - CS 160 fall 2012. https://bid.berkeley.edu/cs160-fall12/index.php/Human_Information_Processing.

Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P. (2015). Role of different factors in predicting movie success. In *2015 international conference on pervasive computing: advance communication technology and application for society* (pp. 1–4). Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.1109/PERVASIVE.2015.7087152.

Burnap, A., Hauser, J. R., & Timoshenko, A. (2019). Design and evaluation of product aesthetics: a human-machine hybrid approach. Available at SSRN 3421771.

Chandrasekaran, D., Srinivasan, R., & Sihi, D. (2017). Effects of offline ad content on online brand search: Insights from super bowl advertising. *Journal of the Academy of Marketing Science*, *46*(3), 403–430. http://dx.doi.org/10.1007/s11747-017-0551-8.

Choi, D., Bang, H., Wojdynski, B. W., Lee, Y. I., & Keib, K. M. (2018). How brand disclosure timing and brand prominence influence consumer's intention to share branded entertainment content. *Journal of Interactive Marketing*, *42*, 18–31. http://dx.doi.org/10.1016/j.intmar.2017.11.001.

Choi, H. J., & Johnson, S. D. (2005). The effect of context-based video instruction on learning and motivation in online courses. *American Journal of Distance Education*, *19*(4), 215–227. http://dx.doi.org/10.1207/s15389286ajde1904_3.

Choudhury, P., Wang, D., Carlson, N. A., & Khanna, T. (2019). Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal*, *40*(11), 1705–1732. http://dx.doi.org/10.1002/smj.3067.

Cisco (2019). Cisco visual networking index: Forecast and trends, 2017–2022 white paper. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html.

Clement, M., Wu, S., & Fischer, M. (2014). Empirical generalizations of demand and supply dynamics for movies. *International Journal of Research in Marketing*, *31*(2), 207–223. http://dx.doi.org/10.1016/j.ijresmar.2013.10.007.

Couwenberg, L. E., Boksem, M. A., Dietvorst, R. C., Worm, L., Verbeke, W. J., & Smidts, A. (2017). Neural responses to functional and experiential ad appeals: Explaining ad effectiveness. *International Journal of Research in Marketing*, *34*(2), 355–366. http://dx.doi.org/10.1016/j.ijresmar.2016.10.005.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In CVPR09.

Dessart, L. (2018). Do ads that tell a story always perform better? The role of character identification and character type in storytelling ads. *International Journal of Research in Marketing*, *35*(2), 289–304. http://dx.doi.org/10.1016/j.ijresmar.2017.12.009.

Dessart, L., & Pitardi, V. (2019). How stories generate consumer engagement: An exploratory study. *Journal of Business Research*, *104*, 183–195. http://dx.doi.org/10.1016/j.jbusres.2019.06.045.

Dhaoui, C., & Webster, C. M. (2020). Brand and consumer engagement behaviors on facebook brand pages: Let's have a (positive) conversation. *International Journal of Research in Marketing*, http://dx.doi.org/10.1016/j.ijresmar.2020.06.005.

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, *53*(6), 881–893. http://dx.doi.org/10.1287/mnsc.1060.0668.

Eliashberg, J., & Sawhney, M. S. (1994). Modeling goes to hollywood: Predicting individual differences in movie enjoyment. *Management Science*, *40*(9), 1151–1173. http://dx.doi.org/10.1287/mnsc.40.9.1151.

Elliot, A., Fairchild, M., & Franklin, A. (2015). Handbook of color psychology. In A. J. Elliot, M. D. Fairchild, & A. Franklin (Eds.), *Cambridge Handbooks in Psychology*, *Handbook of Color Psychology*. Cambridge University Press, http://dx.doi.org/10.1017/cbo9781107337930.

Filntisis, P. P., Efthymiou, N., Koutras, P., Potamianos, G., & Maragos, P. (2019). Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction. *IEEE Robotics and Automation Letters*, *4*(4), 4011–4018. http://dx.doi.org/10.1109/LRA.2019.2930434, arXiv:1901.01805.

Fossen, B. L., & Schweidel, D. A. (2019a). Measuring the impact of product placement with brand-related social media conversations and website traffic. *Marketing Science*, *38*(3), 481–499. http://dx.doi.org/10.1287/mksc.2018.1147.

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, *20*(3–4), 121–136. http://dx.doi.org/10.1007/bf00342633.

Geuens, M., De Pelsmacker, P., & Faseur, T. (2011). Emotional advertising: Revisiting the role of product category. *Journal of Business Research*, *64*(4), 418–426. http://dx.doi.org/10.1016/j.jbusres.2010.03.001.

Goodrich, K., Schiller, S. Z., & Galletta, D. (2015). Consumer reactions to intrusiveness of online-video advertisements: Do length, informativeness, and humor help (or hinder) marketing outcomes? *Journal of Advertising Research*, *55*(1), 37–50. http://dx.doi.org/10.2501/JAR-55-1-037-050.

Guitart, I. A., Gonzalez, J., & Stremersch, S. (2018). Advertising non-premium products as if they were premium: The impact of advertising up on advertising elasticity and brand equity. *International Journal of Research in Marketing*, *35*(3), 471–489. http://dx.doi.org/10.1016/j.ijresmar.2018.03.004.

Gylfe, P., Franck, H., Lebaron, C., & Mantere, S. (2016). Video methods in strategy research: Focusing on embodied cognition. *Strategic Management Journal*, *37*(1), 133–148. http://dx.doi.org/10.1002/smj.2456.

Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2020). The power of brand selfies in consumer-generated brand imagery. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.3354415.

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38. http://dx.doi.org/10.1016/j.ijresmar.2018.09.009.

Hasler, D., & Süsstrunk, S. (2003). Measuring colourfulness in natural images. *Proceedings of SPIE - The International Society for Optical Engineering, Human Vision and Electronic Imaging VIII*, *5007*, 87–95. http://dx.doi.org/10.1117/12.477378.

Hautz, J., Füller, J., Hutter, K., & Thürridl, C. (2014). Let users generate your video ads? The impact of video source and quality on consumers' perceptions and intended behaviors. *Journal of Interactive Marketing*, *28*(1), 1–15. http://dx.doi.org/10.1016/j.intmar.2013.06.003.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hildebrand, C., Efthymiou, F., Busquet, F., Hampton, W. H., Hoffman, D. L., & Novak, T. P. (2020). Voice analytics in business research: conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research*, *121*, 364–374. http://dx.doi.org/10.1016/j.jbusres.2020.09.020.

Himes, S. M., & Thompson, J. K. (2007). Fat stigmatization in television shows and movies: A content analysis. *Obesity*, *15*(3), 712–718. http://dx.doi.org/10.1038/oby.2007.635.

Hui, S. K., Huang, Y., Suher, J., & Inman, J. J. (2013). Deconstructing the "first moment of truth": Understanding unplanned consideration and purchase conversion using in-store video tracking. *Journal of Marketing Research*, *50*(4), 445–462. http://dx.doi.org/10.1509/jmr.12.0065.

Hui, S. K., Meyvis, T., & Assael, H. (2014). Analyzing moment-to-moment data using a Bayesian functional linear model: Application to TV show pilot testing. *Marketing Science*, *33*(2), 222–240. http://dx.doi.org/10.1287/mksc.2013.0835.

Iqbal, S., Qureshi, A. N., & Lodhi, A. M. (2018). Content based video retrieval using convolutional neural network. In *Advances in intelligent systems and computing* (pp. 170–186). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-01054-6_12.

Jeon, Y., Son, H., Chung, A., of Interactive, M. D. J., & 2019, U. (2019). Temporal certainty and skippable in-stream commercials: Effects of ad length, timer, and skip-ad button on irritation and skipping behavior. *Journal of Interactive Marketing*, *47*, 144–158.

Joost, V. D. W., Cordelia, S., Jakob, V., & Diane, L. (2009). Learning color names for real-world applications. *IEEE Transactions on Image Processing*, *18*(7), 1512–1523.

KDnuggets (2017). Most popular language machine learning. https://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html.

Koch, G. R. (2015). Siamese neural networks for one-shot image recognition. In *International conference on machine learning*.

Kretschmer, T., & Peukert, C. (2020). Video killed the radio star? Online music videos and recorded music sales. *Information Systems Research*, http://dx.doi.org/10.1287/isre.2019.0915.

Kumar, A., & Tan, Y. R. (2015). The demand effects of joint product advertising in online videos. *Management Science*, *61*(8), 1921–1937. http://dx.doi.org/10.1287/mnsc.2014.2086.

Lakens, D., Fockenberg, D. A., Lemmens, K. P., Ham, J., & Midden, C. J. (2013). Brightness differences influence the evaluation of affective pictures. *Cognition and Emotion*, *27*(7), 1225–1246. http://dx.doi.org/10.1080/02699931.2013.781501.

Lang, A., Geiger, S., Strickwerda, M., & Sumner, J. (1993). The effects of related and unrelated cuts on television viewers' attention, processing capacity, and memory. *Communication Research*, *20*(1), 4–29. http://dx.doi.org/10.1177/009365093020001001.

Lang, A., Park, B., Sanders-Jackson, A. N., Wilson, B. D., & Wang, Z. (2007). Cognition and emotion in TV message processing: How valence, arousing content, structural complexity, and information density affect the availability of cognitive resources. *Media Psychology*, *10*(3), 317–338. http://dx.doi.org/10.1080/15213260701532880.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. http://dx.doi.org/10.1162/neco.1989.1.4.541.

Leskin, P. (2020). Tiktok surpasses 2 billion downloads and sets a record for app installs in a single quarter. https://www.businessinsider.com/tiktok-app-2-billion-downloads-record-setting-q1-sensor-tower-2020-4?r=DE&IR=T.

Li, X., Shi, M., & Wang, X. S. (2019). Video mining: Measuring visual information using automatic methods. *International Journal of Research in Marketing*, *36*(2), 216–231. http://dx.doi.org/10.1016/j.ijresmar.2019.02.004.

Li, Y., & Xie, Y. (2019). Is a picture worth a thousand words? An empirical study of imagery content and social media engagement. *Journal of Marketing Research*, *57*(1), 1–19. http://dx.doi.org/10.1177/0022243719881113.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science*: *Vol. 8693*, *Computer Vision – ECCV 2014* (pp. 740–755). http://dx.doi.org/10.1007/978-3-319-10602-1_48, arXiv:1405.0312.

Liu, X., Shi, S. W., Teixeira, T., & Wedel, M. (2018). Video content marketing: The making of clips. *Journal of Marketing*, *82*(4), 86–101. http://dx.doi.org/10.1509/jm.16.0048.

Liu-Thompkins, Y., & Rogerson, M. (2012). Rising to stardom: An empirical investigation of the diffusion of user-generated content. *Journal of Interactive Marketing*, *26*(2), 71–82. http://dx.doi.org/10.1016/j.intmar.2011.11.003.

Loewenstein, J., Raghunathan, R., & Heath, C. (2011). The repetition-break plot structure makes effective television advertisements. *Journal of Marketing*, *75*(5), 105–119. http://dx.doi.org/10.1509/jmkg.75.5.105.

Lu, S., Xiao, L., & Ding, M. (2016). A video-based automated recommender (VAR) system for garments. *Marketing Science*, *35*(3), 484–510. http://dx.doi.org/10.1287/mksc.2016.0984.

Marinova, D., Singh, S. K., & Singh, J. (2018). Frontline problem-solving effectiveness: A dynamic analysis of verbal and nonverbal cues. *Journal of Marketing Research*, *55*(2), 178–192. http://dx.doi.org/10.1509/jmr.15.0243.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of workshop At ICLR*.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543. http://dx.doi.org/10.1287/mksc.1120.0713.

Pech-Pacheco, J. L., Cristóbal, G., Chamorro-Martínez, J., & Fernández-Valdivia, J. (2000). Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th international conference on pattern recognition: Vol. 3* (pp. 314–317).

Quesenberry, K. A., & Coolsen, M. K. (2019). Drama goes viral: Effects of story development on shares and views of online advertising videos. *Journal of Interactive Marketing*, *48*, 1–16. http://dx.doi.org/10.1016/j.intmar.2019.05.001.

van Reijmersdal, E. A., Rozendaal, E., Hudders, L., Vanwesenbeeck, I., Cauberghe, V., & van Berlo, Z. M. (2020). Effects of disclosing influencer marketing in videos: An eye tracking study among children in early adolescence. *Journal of Interactive Marketing*, *49*, 94–106. http://dx.doi.org/10.1016/j.intmar.2019.09.001.

Roberts, K., Roberts, J. H., Danaher, P. J., & Raghavan, R. (2015). Incorporating emotions into evaluation and choice models: Application to kmart Australia. *Marketing Science*, *34*(6), 815–824. http://dx.doi.org/10.1287/mksc.2015.0954.

Schikowsky, A., Hartmann, J., Heitmann, M., & Haenlein, M. (2020). *Mining iconic marketing assets: A unified multi-modal deep learning framework*. Working paper.

Schindler, P. S. (1986). Color and contrast in magazine advertising. *Psychology and Marketing*, *3*(2), 69–78. http://dx.doi.org/10.1002/mar.4220030203.

Shehu, E., Bijmolt, T. H., & Clement, M. (2016). Effects of likeability dynamics on consumers' intention to share online video advertisements. *Journal of Interactive Marketing*, *35*, 27–43. http://dx.doi.org/10.1016/j.intmar.2016.01.001.

Shou, Z., Pan, J., Chan, J., Miyazawa, K., Mansour, H., Vetro, A., Giro-I-Nieto, X., & Chang, S. F. (2018). Online detection of action start in untrimmed, streaming videos. In *LNCS*: *Vol. 11207*, *The european conference on computer vision* (pp. 534–551). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-01219-9_33, arXiv:1802.06822.

Simmonds, L., Bellman, S., Kennedy, R., Nenycz-Thiel, M., & Bogomolova, S. (2019). Moderating effects of prior brand usage on visual attention to video advertising and recall: An eye-tracking investigation. *Journal of Business Research*, *111*, 241–248. http://dx.doi.org/10.1016/j.jbusres.2019.02.062.

Smith, R. (2007). An overview of the tesseract OCR engine. In *Ninth international conference on document analysis and recognition* (pp. 629–633).

Somers, J. (2018). The scientific paper is obsolete. here's what's next. - the atlantic. https://www.theatlantic.com/science/archive/2018/04/the-scientific-paper-is-obsolete/556676/.

Statista (2019). Digital advertising report 2019 - video advertising. https://de.statista.com/statistik/studie/id/41114/dokument/digital-advertising-report-video-advertising/.

Swani, K., & Milne, G. R. (2017). Evaluating facebook brand content popularity for service versus goods offerings. *Journal of Business Research*, *79*, 123–133. http://dx.doi.org/10.1016/j.jbusres.2017.06.003.

Teixeira, T. S., Wedel, M., & Pieters, R. (2010). Moment-to-moment optimal branding in TV commercials: Preventing avoidance by pulsing. *Marketing Science*, *29*(5), 783–804. http://dx.doi.org/10.1287/mksc.1100.0567.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., & Zhang, Y. (2019). What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence. *Journal of Marketing*, *83*(4), 1–20. http://dx.doi.org/10.1177/0022242919841034.

Tomalski, P., Csibra, G., & Johnson, M. H. (2009). Rapid orienting toward face-like stimuli with gaze-relevant contrast information. *Perception*, *38*(4), 569–578. http://dx.doi.org/10.1068/p6137.

Tucker, C. E. (2015). The reach and persuasiveness of viral video ads. *Marketing Science*, *34*(2), 281–296. http://dx.doi.org/10.1287/mksc.2014.0874.

Vijayakumar, V., & Nedunchezhian, R. (2012). A study on video data mining. *International Journal of Multimedia Information Retrieval*, *1*(3), 153–172. http://dx.doi.org/10.1007/s13735-012-0016-2.

Wood, R. T. A., Griffiths, M. D., Chappell, D., & Davies, M. N. O. (2004). The structural characteristics of video games: A psycho-structural analysis. *CyberPsychology & Behavior*, *7*(1), 1–10.

Xiao, L., & Ding, M. (2014). Just the faces: Exploring the effects of facial features in print advertising. *Marketing Science*, *33*(3), 338–352. http://dx.doi.org/10.1287/mksc.2013.0837.

Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 435–442).

Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2017). How much is an image worth? Airbnb property demand estimation leveraging large scale image analytics. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.2976021.

Zhang, X., Li, S., Burke, R. R., & Leykin, A. (2014). An examination of social influence on shopper behavior using video tracking data. *Journal of Marketing*, *78*(5), 24–41. http://dx.doi.org/10.1509/jm.12.0106.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. http://dx.doi.org/10.1109/LSP.2016.2603342, arXiv:1604.02878.

Jasper Schwenzow is a doctoral student at Hamburg University. His substantive research interests include visual communication, social media marketing and customer insights. He explores consumer psychology at scale by leveraging and extending state-of-the-art technologies in the field of computer vision and web scraping. He is part of the research unit "Marketing of Products in the Age of Digital Social Media" (DFG-FG 1452). Before joining Hamburg University, he worked as a management consultant at McKinsey & Company.