applied sciences

MDPI

Article

Low-Level Video Features as Predictors of Consumer Engagement in Multimedia Advertisement

Evin Aslan Oğuz [1,2,†], Andrej Košir [1,†], Gregor Strle [1,3,†] and Urban Burnik [1,*,†]

[1] User-Adapted Communication and Ambient Intelligence Lab, Faculty of Electrical Engineering, University of Ljubljana, SI 1000 Ljubljana, Slovenia
[2] Nielsen Lab d.o.o., SI 1000 Ljubljana, Slovenia
[3] Scientific Research Centre, ZRC SAZU, SI 1000 Ljubljana, Slovenia
* Correspondence: urban.burnik@fe.uni-lj.si; Tel.: +386-1-476-8440
† These authors contributed equally to this work.

Abstract: The article addresses modelling of consumer engagement in video advertising based on automatically derived low-level video features. The focus is on a young consumer group (18–24 years old) that uses ad-supported online streaming more than any other group. The reference ground truth for consumer engagement was collected in an online crowdsourcing study (N = 150 participants) using the User Engagement Scale-Short Form (UES-SF). Several aspects of consumer engagement were modeled: focused attention, aesthetic appeal, perceived usability, and reward. The contribution of low-level video features was assessed using both the linear and nonlinear models. The best predictions were obtained for the UES-SF dimension Aesthetic Appeal ($R^2 = 0.35$) using a nonlinear model. Overall, the results show that several video features are statistically significant in predicting consumer engagement with an ad. We have identified linear relations with Lighting Key and quadratic relations with Color Variance and Motion features ($p < 0.02$). However, their explained variance is relatively low (up to 25%).

Keywords: multimedia signal processing; low-level video features; consumer engagement; automatic estimation; advertising exposure

check for updates

Citation: Oğuz, E.A.; Košir, A.; Strle, G.; Burnik, U. Low-Level Video Features as Predictors of Consumer Engagement in Multimedia Advertisement. *Appl. Sci.* **2023**, *13*, 2426. https://doi.org/10.3390/app13042426

Academic Editor: Andrea Prati

Received: 21 November 2022
Revised: 2 February 2023
Accepted: 9 February 2023
Published: 13 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid and sustained development of new content consumption models based on personalized media access and streaming is having a significant impact on the established advertising landscape [1]. Audience measurement providers are struggling to keep pace with these advances and to provide advertisers with improved advertising effectiveness metrics. The latter requires robust metrics to demonstrate the impact of advertising campaigns on potential consumers.

Exposure to ads can be quantified in the context of consumer engagement (UE). To this end, brands use innovative tactics to create memorable impressions in order to increase consumer awareness and engagement with ads. Higher engagement leads to more ad exposure, improves brand memorability, and positively affects consumer purchase intentions [2–4]. However, exposure to ads and its effects are complex and depend on several factors. Subjective and objective factors that affect ad exposure include the circumstances in which a particular advertisement is consumed (context, denoted by c), subjective conditions related to a consumer's personality and/or temporal mood (person, denoted by p), and the characteristics of the advertising video materials (video, denoted by v). A measured engagement using of UES-SF instrument is denoted by UE, and the measurement noise of this measurement is denoted by ε:

$$f(c, p, v) = UE + \varepsilon \quad (1)$$

A thorough model of ad engagement necessarily considers all of the three above mentioned factor groups. The present research focuses on parts related to the impact of video material characteristics on user engagement. We present a novel approach to modeling the relation between consumer ad engagement and visual video features derived from ads. The advantage of the proposed approach is that it is unobtrusive, scalable, and can be automated. To our knowledge, most of the existing approaches estimate consumer engagement based on user and content metadata (such as the number of views, video resolution, and genre), ignoring the potential effects that visual video features might have on consumer engagement. In addition, many existing approaches lack a referential ground truth for user-perceived engagement. We hypothesize that the low-level video features of an ad might capture some of its visual effects on consumers that cannot be modeled from the metadata itself.

The article has two objectives: a) to identify relevant low-level video features that can be automatically derived from video ad, and b) to test the feasibility of the model built on these features by evaluating their predictive power for assessing consumer engagement. The reference ground truth for the consumer engagement was obtained using the User Engagement Scale-Short Form (UES-SF), a standardized psychometric instrument which has been shown to capture several aspects of engagement (focused attention, aesthetic appeal, perceived usability, and reward) and provide reliable results [5]. The ground truth was collected in an online crowdsourcing study (N = 150 participants) that measured participants' engagement in video ads using UES-SF. The target audience of the present study was young adults, the largest segment of digital natives who engage in online multimedia, especially ad-supported video streaming. The focus was on short-term exposure to in-video ads, which are common in ad-supported video streaming [6].

The remainder of this paper is organized as follows: Section 2 briefly introduces related work and discusses the concepts of ad exposure, consumer engagement, and relevant low-level video features. Section 3 presents the materials and methods used in the observational study, gathering the ground truth of perceived consumer engagement, and constructing a model based on video features. Section 4 presents the results of the study and the evaluation of low-level video features for assessing consumer ad engagement scores. In Section 5, we discuss our findings, shortcomings of the presented research, and future work.

2. Related Work

2.1. Ad Exposure Effects and Consumer Engagement

Consumer engagement is a multidimensional concept that encompasses the cognitive, psychological, and behavioral components of consumer–brand interaction [2–4,7,8]. It is inextricably linked to ad exposure and is the main factor in its effectiveness [2,4]. Consumer engagement is key to understanding consumer behavior, along with memorability [9], attitude [10], and purchase intention [4]. Higher consumer engagement leads to more exposure, improves brand memorability and familiarity, and may positively affect consumer purchase intentions [2–4]. Therefore, consumer engagement is one of the most important metrics for quantifying ad exposure and its effectiveness, and more generally, brand attitudes and overall consumer behavior.

Consequently, the cognitive, affective, and behavioral dimensions of engagement are increasingly considered in the digital advertising landscape [2,4]. However, despite the growing trend toward measuring engagement, the field lacks unifying theories and models that provide empirical support for consumer ad engagement [4,11]. An analysis of over 200 studies on media exposure conducted by [11] found that there is still no universally accepted conceptualization and operationalization of media exposure and its effects, including consumer engagement. As [2] point out, there is no one-size-fits-all solution because consumer engagement is based on context-specific experiences and settings.

For example, some researchers operationalize engagement in terms of behavioral manifestations in the consumer–brand relationship [8]. Others consider engagement in terms of brand messages evoking passionate and affective responses in consumers [7] or enabling

creative experiences [9]. The cognitive and emotional dimensions of consumer engagement were conceptualized in [12,13]. For example, Ref. [13] conceptualized engagement in the context of emotional arousal and created a model based on patterns of physiological responses to multimedia stimuli. A more complex conceptualization of consumer engagement was provided by [12], who defined the consumer engagement framework as a multidimensional construct based on cognitive, emotional, experiential, and social dimensions. The behavioral and social dimensions of engagement were further conceptualized using the User Engagement Scale (UES) developed by [5]. The UES measures the subjective aspects of engagement across multiple dimensions: aesthetic appeal, focused attention, novelty, perceived usability, felt involvement, and endurability [5].

The above operationalizations of engagement depend on user feedback from user surveys, interviews, or self-reports. However, it is not always possible to collect user feedback, which is time-consuming, and self-reports have been shown to be unreliable, as they are prone to rater bias [14]. This is especially true for digital advertising services, such as ad-supported video streaming, where traditional methods for gathering user feedback are difficult, if not impossible, to implement [6,10]. More direct methods for measuring consumer engagement are required to capture consumer engagement in real time, without interrupting the streaming experience.

To this end, several novel approaches have been developed to directly predict engagement based on machine learning and feature modeling of the content itself. For example, Ref. [15] proposed a graph-embedding model called DeepWalk to predict the video engagement of an ad to detect ad fraud. This model can detect fake videos and fraud patterns in videos for well-known brands. More generally, Ref. [16] proposed an automatic approach for processing and evaluating learner engagement. They developed a prediction model for context-agnostic engagement based on the video features of learning content and engagement signals. Research conducted by [17] on YouTube review videos (totaling 600 h) identified features and patterns relevant to emotion (valence and arousal) and trustworthiness as the underlying dimensions of engagement. Several indicators of user engagement, including views, the ratio of likes to dislikes, and sentiment in comments, served as the ground truth. Ref. [18] defined a set of video metrics (including video quality, time, and average percentage of video viewed) to model relative engagement, based on 5.3 million YouTube videos. Their results show that video engagement metrics are stable over time, with most of their variance explained by video context, topics, and channel information. Moreover, the study also found that the time spent watching is a better indicator of engagement than the number of views statistics, which is widely used in current studies of ad engagement.

The advantage of these modeling approaches is that they are scalable, automatic, and, in several cases, can directly measure consumer engagement. However, developing robust predictive models for consumer engagement remains a challenge. Most of these approaches rely on user metadata (e.g., view count, frequency, and likeability data) or video quality features (e.g., video resolution), and lack the reference ground truth of user-perceived engagement. A summary of these approaches is collected in Table 1. To the best of our knowledge, none of the existing approaches use features derived from the video itself, which is the goal of the present research.

The video and multimedia descriptors are presented in the following sections. These descriptors were later used in the proposed model of consumer ad engagement, along with the referential data of perceived consumer engagement obtained from the observational study.

Table 1. Contributions of related work on consumer engagement.

Research Task	Method	Findings
Measuring engagement based on affective and social cues [7]	A survey and literature review based on affective and social cues of engagement	Conceptualisation of engagement based on various ways the engagement information can be collected and used. Focus is on the affective cues, social cues, and interactive media.
Conceptualization of consumer brand engagement [12]	Conceptualization of CBE based on literature review and advertising practitioners' perceptions of CBE. Focus is on cognitive, affective, conative, experimental, and social cues.	CBE "emerges as a multi-dimensional construct that beyond traditional cognitive, emotional and conative dimensions seems to be based on emerging experiential and social dimensions that appear as its central elements."
Development of CBE scale [8]	Conceptualisation and operationalisation of CBE scale, focusing on brand engagement and social media advertising.	CBE scale: Consumer Brand Engagement Scale based on cognitive, affective, and activation cues.
Development of engagement scale (brand/community) [3]	Conceptualisation and operationalisation of consumer engagement in the context of online brand communities	A 22-item scale of consumer engagement focused on brand and community.
Conceptualization of a flexible approach to measuring engagement [2]	A survey and literature review	A flexible approach to measuring engagement is needed, taking into account context-specific experiences that can vary across brands and products.
Conceptualization of consumer engagement in digital ecosystem and computational advertising [4]	A survey and literature review	Conceptualisation of consumer engagement through meaningful and sustained interactions with consumers in digital ecosystems and computational advertising.
Research designs that support repeated measures of consumer engagement [10]	A survey and literature review	Research designs that support repeated observations of individual attitudes and behavior should be well positioned to study media effects and processes in the digital age.
Memory performance in ad exposure [9]	Recognition tests using retrospective confidence judgments	Exploratory study measuring the effects of creative ads and insight on memory performance. The findings suggest that insight makes advertisements more memorable, especially those that are creative.
Prediction of engagement based on physiological responses to video ads using machine learning [13]	Classification of emotional arousal based on affect and physiological responses (GSR, ECG, EOG, EEG, and PPG).	Methods of affective computing, emotion recognition, classification were used. The study found that the patterns of physiological response to each multimedia stimuli were effective in the classification of stimulus type.
Development of user engagement scale [5]	Validation of UES-SF: User Engagement Scale	A verified a four-factor structure (aesthetic appeal, focused attention, perceived usability, and reward) for the 12-item of User Engagement Scale short form was tested in various digital scenarios (e.g., e-shopping).
Prediction of video engagement using machine learning [18]	Prediction of video engagement based on a large-scale measurement of video engagement from a collection of 5.3 million YouTube videos.	The study proposes a new metric, relative engagement that is calibrated against video properties and strongly correlates with recognized notions of quality.
Prediction of video engagement using machine learning [15]	A graph-embedding model to predict video engagement of an advertisement based on viewing patterns over time, using only a few samples.	The learned embedding is able to identify viewing patterns of fraud and popular videos.
Prediction of video engagement using machine learning [17]	Prediction of user engagement indicators based on the interpretable patterns of arousal, valence, and trustworthiness. The analysis was conducted on the emotion-annotated dataset of 600 YouTube videos.	Smaller boundary ranges and fluctuations for arousal lead to an increase in user engagement. Furthermore, the time-series features reveal significant correlations for each dimension, such as count below signal mean (arousal), number of peaks (valence), and absolute energy (trustworthiness). From this, an effective combination of features is outlined for approaches aiming to automatically predict several user engagement indicators.

2.2. Video and Multimedia Content Descriptors

The problem of quantitatively assessing the vast amounts of data in modern multimedia content has been addressed for decades, first for the purpose of content indexing and retrieval and later for content-based recommender systems (CBRS). We hypothesize that similar dimensionality reduction tools originally used for indexing and CBRS can also be used to determine the impact of content on multimedia engagement scales.

Many user-based multimedia classification systems utilize high-level content features, such as genre, director, cast, tags, text reviews, and similar meta-information [19]. These metadata annotations are scarce, particularly for video advertising, and implicit metadata available through computational analysis of the content are preferred. Therefore, the main goal of the presented research is to relate an individual's measurable content engagement to the computable features of the multimedia content consumed. We begin our presentation of potential feature sets with a standardized and widely used set of low-level descriptors for visual multimedia content, a subset of which can be algorithmically extracted.

2.2.1. MPEG-7

The MPEG-7 represents a well-established set of semantic and descriptive features that help manage, search, analyze, and understand modern multimedia content. Multimedia in the form of large numeric datasets poses a challenge to their interpretation due to the lack of efficient identification techniques that deal with large datasets at the specific level of content description defined by the context of a current application. A higher-level, yet generic set of media record descriptors that would reduce the size of the original content notation would be helpful. These descriptors can be determined manually or by automatic content analysis and preprocessing. The established MPEG7 [20,21] standard addresses generic media record descriptors for indexing and searching multimedia databases, and has traditionally been used as a tool for indexing, searching, and accessing the desired multimedia content.

MPEG-7 comprises a family of standards for multimedia content description and is defined by ISO / IEC 15938 [20]. Unlike its predecessors, the standard is not intended to encode audiovisual material but to describe it. Metadata about the material is stored in XML format and provides an accurate description that can be used to mark individual segments related to events in the material with a time code. This standard was developed to allow users to quickly and efficiently retrieve multimedia materials they need. It can be used independently of other MPEG standards and is closely related to the nature of the audiovisual material, as defined in MPEG-7. The MPEG-7 architecture states that multimedia content descriptions must be kept separate and the relationships between them must be unambiguously preserved. MPEG-7 is based on four basic building blocks: *Descriptors* (metadata extracted from multimedia materials), *Description Schemes* (structures and semantics of description relationships), *Description Definition Language (DDL)* (a language that allows you to create and add new description schemas and descriptions), and *System Tools* (tools that support description multiplexing, content synchronization, downloading, record encoding, and intellectual property protection).

MPEG-7 contains 13 volumes, of which Part 3 contains tools for visual content description, namely, static images, videos, and 3D models. Descriptions of the visual features defined by the standard include color, texture, shape, and motion, as well as object localization and face recognition, each of which is multi-dimensional. Specifically, the core MPEG-7 visual descriptors are Color Layout, Color Structure, Dominant Color, Scalable Color (color), Edge Histogram, Homogeneous Texture, Texture Browsing (texture), Region-based Shape, Contour-based Shape (shape), Camera Motion, Parametric Motion, and Motion Activity (motion). Other descriptors represent structures for aggregation and localization and include core descriptor aggregates with additional semantic information. These include Scalable Color descriptions (Group-of-Frames/Group-of-Pictures), 3D mesh information (Shape 3D), object segmentation in subsequent frames (Motion Trajectory) and specific face extraction results (Face Recognition), Spatial 2D Coordinates, Grid Layout, Region

Locator, Time Series, Temporal Interpolation and Spatio-Temporal Locator. The remaining set of aggregated descriptors contains supplementary (textual) structures in color spaces, color quantization, and multiple 2D views of 3D objects. Some of these descriptors can be obtained using automated image analysis [22].

The visual features of MPEG-7 have relatively high dimensionality because each feature is a vector. Studies also demonstrated a high degree of redundancy in MPEG-7 visual features [23]. The MPEG-7 features are standardized, well-structured, computationally available, and provide a good basis for indexing, searching, and querying multimedia content. Some key computable features are represented in great detail (e.g., color and texture), whereas other media aesthetic elements are not directly covered. However, there is no commonly known direct interpretation of MPEG-7 visual features in terms of arousal triggered by media consumers.

2.2.2. Stylistic Video Features

Research on applied media aesthetics demonstrates that the effects of production techniques on viewers' aesthetic reflexes can be evaluated computationally, based on the aesthetic elements of multimedia items [24]. For example, [24] described how the basic elements of media aesthetics can be used to produce maximally effective visual appeal of images and videos. Another study on multimedia services demonstrated filtering of large information spaces to classify the attractiveness of multimedia items to users [25]. Other researchers have recommended the use of automatically extracted low-level features in hybrid and autonomous recommender systems [26,27]. We hypothesize that the knowledge of individuals' subjective responses to multimedia content, based on low-level features from applied media aesthetics [24] and recommender systems [26,27] can be extended to multimedia exposure.

The elements of media aesthetics refer to the use of light, colors, screen area, depth, and volume in the three-dimensional realm, time, and movement in videos. The expressive communication of movie producers is based on established design rules, also known as movie grammar. Ref. [28] identified the elements of movie grammar using computable video features to identify the movie genre based on video previews. Their proposal included lightning key, color variance, motion content, and average shot length features along with a well-described algorithm for determining each feature. An improved proposal for aesthetically inspired descriptors applied to a movie recommendation system was presented by Deldjoo et al. [26] and is called *stylistic video features*. We adopted a basic set of visual features related to color, illuminance, object motion, and shot tempo of the video sequence, as described below.

Lighting Key is a powerful tool used by directors to control the emotions evoked by video consumers [26,28]. The use of *high-key lighting*, where less contrast is used, resulting in bright images, promotes positive and optimistic responses. The use of dark shading, resulting in low mean and standard deviation grayscale, creates a mysterious and dramatic mood typical of noir movies; this is known as *low-key lighting*. The Lighting Key ξ_q of a frame is defined as the product of the mean μ and standard deviation σ of the brightness component (V) in the three-dimensional color space HSV,

$$\xi_q = \mu \cdot \sigma. \tag{2}$$

Typically, high values of both the mean and standard deviation of brightness are associated with funny video scenes, which may be associated with $\xi_q > \tau_c$. On the other hand, dark, dramatic scenes indicate the opposite situation, where $\xi_q < \tau_h$. In some video materials, where $\tau_h < \xi_q < \tau_c$ lighting key is less pronounced, which directly affects viewer mood [26].

The Lighting Key value for an entire video is defined as the average of ξ_q. To speed up the calculations, we can take the middle frame of each shot as a representative value, resulting in

$$\mu_{lk} = \frac{\sum_{q=1}^{n_{sh}} \xi_q}{n_{sh}}. \tag{3}$$

Color Variance is another indicator closely related to the mood that a director intends to express with the video [26,28]. Color Variance is a scalar dimension that indicates how vibrant a movie's color palette is. Lighter colors are generally used in light and funny videos, whereas darker colors are predominant in horror and morbid scenes.

We can calculate the Color Variance in the three-dimensional color space Luv using the covariance matrix of the three color dimensions. The value of the Color Variance per frame is equal to

$$\Sigma_q = \begin{vmatrix} \sigma_L^2 & \sigma_{Lu}^2 & \sigma_{Lv}^2 \\ \sigma_{Lu}^2 & \sigma_u^2 & \sigma_{uv}^2 \\ \sigma_{Lv}^2 & \sigma_{uv}^2 & \sigma_v^2 \end{vmatrix} \tag{4}$$

The Final Color Variance value for the entire video is the average of Σ_q using the representative key frames of each shot, resulting in

$$\mu_{cv} = \frac{\sum_{q=1}^{n_{sh}} \Sigma_q}{n_{sh}} \tag{5}$$

Motion Activity within a video refers to movements of the camera or objects within the scene. The shot and cut dynamics are also included to a certain degree in the Average Shot Length video features described in the next subsection. However, object dynamics remain an independent feature related to viewers' feelings evoked by videos. Some early studies used MPEG motion vectors in the compression domain to obtain motion information of a video [29]. Ref. [28] uses a method based on *visual disturbance* of the image sequence. Ref. [26] proposes to characterize motion activity using optical flow [30], which indicates the velocity of objects in a sequence of images. According to their proposal, both the mean value of the pixel motions

$$\mu_{\overline{m}} = \frac{\sum_{t=1}^{n_f} \overline{m}_t}{n_f} \tag{6}$$

and standard deviation of pixel motions

$$\mu_{\sigma_m^2} = \frac{\sum_{t=1}^{n_f} (\sigma_m^2)_t}{n_f} \tag{7}$$

can be applied as motion features of a video.

Average Shot Length is defined as the average of the shot duration

$$\overline{L_{sh}} = \frac{n_f}{n_{sh}}, \tag{8}$$

where n_f is the number of frames and n_{sh} is the number of shots in the entire video being evaluated. This ratio provides a good estimate of the tempo of the scenes, which is controlled by the video director. This feature was introduced and evaluated in depth by [31] and was later used by [26,28]. There are many algorithms for automatic scene transition detection that are well described in [28,31]. Well-known techniques for scene detection include content-based, adaptive-, and threshold-based scene detectors. These three techniques are used in an open-source scene detection tool *PySceneDetect* hosted on GitHub [32]. The content-aware method detects jump cuts within a video and is triggered when the difference between two consecutive frames exceeds a predefined threshold. It operates in the HSV color space across all three color channels. The adaptive method works

in a similar manner but uses the moving average of several adjacent changes to reduce the detection of false scene transitions in the case of fast camera movements. The threshold method is the most traditional method which, according to PySceneDetect, triggers a scene cut when the brightness of an image falls below a predefined threshold (fade-out).

The five presented video features represent a good estimate of the stylistic video features and together they form a video feature vector

$$\mathbf{f_v} = (\mu_{lk}, \mu_{cv}, \overline{L_{sh}}, \mu_{\overline{m}}, \mu_{\sigma_m^2}) \tag{9}$$

comprised of Lighting Key, Color Variance, Average Shot Length, Mean Motion Activity and Standard Deviation of Motion Activity.

At this early stage of development, we decided to use more explanatory and compact stylistic video features rather than MPEG-7. The above feature vector $\mathbf{f_v}$ was therefore used in our study. A collection of related work findings is assembled in Table 2.

Table 2. Contributions of related work regarding the applications of computable low-level video features.

Research Task	Method	Findings
Video indexing and retrieval [22].	Use MPEG-7 Visual features to search and retrieve videos based on distance measures in feature space.	Performance depends on specific case. Other distance measures may be required.
Statistical analysis of MPEG-7 descriptors [23].	Mean and variance of description elements, distribution of elements, cluster analysis (hierarchical and topological) and factor analysis were performed.	The analysis revealed that most MPEG-7 descriptions are highly redundant and sensitive.
Movie recommendation based on MPEG-7 visual features [27].	Deep-learning based recommender system based on MPEG-7 features.	A working proof of concept demonstrates a potential of computable visual features for recommender systems. A large dataset is required for deep neural net training.
Applied media aesthetics [24].	Computational evaluation of production techniques on viewers' aesthetic reflexes based on the aesthetic elements of multimedia items.	Fundamental elements of media aesthetics can be used to evaluate the effective visual appeal of images and videos.
Content-based video recommendation system based on stylistic visual features [26].	Recommender system based on low-level video features defined upon applied media aesthetic elements as a substitute to semantic media metadata.	Compared with semantic content-based recommender systems, the proposed method leads to more accurate results.
The use of computable video features for film classification [28]	Classification of movie genres based on computable video features.	Average Shot Length, Color variance, Lighting Key, and Motion Content features can be used as classifiers of movie genre.

2.3. Research Gap

A review of the current state-of-the-art shows several initiatives in investigating consumer engagement in advertising. However, these efforts are still mainly aimed at conceptualizing engagement towards a unified theory, while operationalisations of engagement into valid psychometric instruments are still relatively rare. Table 1 provides an overview of the different approaches. To our knowledge, there are currently only two standardized instruments for measuring consumer engagement and providing ground truth: Consumer Brand Engagement Scale (CBS) [8] and User Engagement Scale (UES) [5]. The CBS focuses on brand awareness as a particular aspect of consumer engagement, while the UES is more general and robust. Of the applied approaches presented (cf. [9,13,15,17,18]), none uses any of these tools in their modeling, instead relying on combinations of engagement indicators, such as the number of views, the like/dislike ratio, or the sentiment of comments.

Moreover, as can be seen from Table 2, there are no engagement studies based on computable low-level features of the advertising video materials v. The table shows applications of low-level video features relevant to the presented research. MPEG-7 Visual, for example, uses multidimensional features that have proven useful in multimedia indexing

and retrieval [22]. Given the high dimensionality and redundancies found in MPEG-7 descriptors [23], a much larger dataset is required for a modeling attempt using deep learning techniques. Other research shows that visual features related to applied media aesthetics can behave well as media classifiers even for feature vectors with low dimensionality.

3. Materials and Methods

3.1. Video Ads

The study was conducted using video ads. The selection criteria were developed in collaboration with Nielsen Company's media and marketing experts, taking into account the target audience and the different contexts of potential consumer engagement. The focus was on short-term ad exposure, which is common in ad-supported video streaming [6].

The exclusion criteria for the video ads were: (a) content that depicted or promoted smoking, alcohol, or sex; (b) ads that did not promote an actual product but conveyed a social message (HIV testing, no drinking and driving) or promoted a service (e.g., dry cleaning, real estate companies, etc.); (c) low-resolution videos or very old video ads (produced before 1980); or (d) ads that promoted baby- or child-specific products because they were not relevant to the target audience. The inclusion criteria for the video ads were as follows: (a) language (only ads in English), (b) length (between 15 and 75 s in duration), and (c) accessibility (available online).

Based on these criteria, 30 video ads were selected from YouTube. A complete list of videos with access links is provided in the referenced dataset. These video ads were later used to derive low-level video features relevant to modeling of consumer ad engagement.

3.2. Low-Level Video Feature Selection

Based on the selection of video descriptors described in Section 2, a set of low-level video features was obtained from the six video ads used in the observational crowdsourcing study. These video features include Lighting Key, Color Variance, Mean Motion Activity, Standard Deviation of Motion Activity, and Average Shot Length.

For the Lighting Key and Color Variance features, the values for each video frame were determined first. The average of the values assigned to the frames was calculated to obtain a single scalar value for each video. Similarly, optical flow vectors were calculated for each successive pair of video frames. The measurement of motion activity with respect to the two frames was calculated using the mean and standard deviation of the absolute values of the flow vectors. The two corresponding scalar motion-related video features associated with the entire video were then calculated as the average over the corresponding video frames.

The dynamics of the video processing were also considered. The Average Shot Length represents the average video scene length between two video edits during the video production. Scene transition estimates were obtained using the PySceneDetect [32] software package. A content-based detector is also used. This method is based on jump-cut detection and seeks the differences in consecutive video frames exceeding a certain threshold. The experimentally determined threshold was set as $t = 10$, where the threshold was an 8-bit integer. The entire process was set to operate in the HSV color space by using the average value of the three color channels. The scene transitions obtained were thoroughly verified by adjusting the values in rare cases where the automatically determined timestamps of the transitions did not match the actual shot boundaries.

3.3. Referential Ground Truth for Perceived Ad Engagement

The referential ground truth for modeling consumers' perceived ad engagement was collected using the User Engagement Scale-Short Form (UES-SF), which has been shown to provide reliable results [5]. UES-SF is a 12-item psychometric instrument that represents the cognitive, behavioral, and social aspects of engagement in four core components (subscales): Focused Attention, Aesthetic Appeal, Perceived Usability, and Reward. It was developed

as a short and quick alternative to the 31-item UES and is particularly suitable for the online collection of participants' responses [5].

The UES-SF items are rated on a five-point rating scale. The wording of the item questions can be adapted to the specific use case. For example, instructions to participants can be worded as follows, "The following questions aim to rate your experience viewing the advertisement. For each question, please use the following scale to indicate which most closely applies to you." An example of a question item from the Focused Attention component: *"I was absorbed in this advertisement"*.

Scores can be provided for individual components (FA, PU, AE, and RW) or for the entire UES-SF. Scores were calculated as the mean for each component and for each participant by calculating the average of the summed scores for the items in each component. The total UES-SF score per participant was calculated by adding the mean scores of the components. A detailed description of the UES components and items, as well as the scoring procedure, can be found in [5]. The scoring allows for modifications of questions in order to fit the evaluated content type. The exact questions used in the research are part of the referenced and published dataset.

3.4. Observational Study

An online observational study was conducted using the Clickworker crowdsourcing platform (https://www.clickworker.com/, accessed on 20 November 2022) to collect data on consumer engagement with ads. The observational study is part of a broader investigation of consumer behavior related to in-video advertising. It covers multiple aspects of ad exposure, including engagement, attitude, and purchase intent, with psychometric measures of participants' mood, affective state, preferences, and demographics. Consequently, the number of materials used must be minimized to accommodate the scope of the study. In this article, the focus is on ad engagement.

The target population was young adults (18–24 years old), native English speakers from the USA who engage in online multimedia more than any other age group, particularly ad-supported video streaming. 54% of the respondents were female and 46% male. 61% of them were white Caucasian, 8% Hispanic, 16% African American, 2% Native American, 10% Asian and 3% of other ethnicity. In addition, 70% were living in metropolitan and 30% in rural areas. The highest educational degree achieved was 2% primary, 50% secondary, 7% vocational, 35% bachelor's and 6% master's education. Furthermore, 47% of the respondents were students, 23% self-employed, 19% regularly employed and 11% unemployed. The reference ground truth for consumer engagement with ads was obtained using (UES-SF) and includes several aspects of engagement: *focused attention*, *aesthetic appeal*, *perceived usability*, and *reward* [5]. Using UES-SF, each participant rated only one video ad randomly assigned to them. The number of raters per individual video was balanced ($N_{rpv} = 5$). Responses were collected from a total of ($N = 150$) participants.

To control for the technology-related effects of multimedia exposure (e.g., screen size and technology-related usage behavior), only participants who used a personal computer were included in the study. This restriction also reduced the influence of ambient distractions in case of mobile media consumption. Participants were rewarded 2.50 Euro each for participating in the study. First, informed consent was obtained from the participants, informing them of the purpose of the study. They were then given a brief description of the purpose and duration of the study (average duration of five minutes). The participants were informed that no time restrictions were imposed on them, and were instructed to set an appropriate volume on their computer to fully experience the video ad. After viewing the ad, each participant was asked to rate their engagement using UES-SF. A content-related verification question has been implemented to make sure media content was indeed consumed by the participant.

3.5. Data Processing

Video features have been determined according to definitions by using own routines in Matlab R2021a. Statistical analysis and modelling were performed in Python 3.8.13 using libraries pydove 0.3.5, seaborn 0.11.2 and numpy 1.21.5. All major scripts together with the corresponding dataset are available on GitHub https://github.com/ub-lucami/llvfeatues, accessed on 21 November 2022.

4. Results

The following sections report the results of the observational study described in Section 3 and of the modeling of consumers' ad engagement.

4.1. Descriptive Statistics for Video Features and Engagement Scores

4.1.1. Video Features

The basic statistics for the video features are shown in Figure 1. The figure includes the minimum, maximum, mean, and standard deviation of the corresponding video features for each ad. Each set is complemented with a randomized bar graph of raw consecutive feature values and histogram plots thereof.

	min	max	mean	std	raw	hist
Lighting Key	0.0572	0.195	0.117	0.04		
Color Variance	7.69×10^{3}	4.043×10^{7}	6.26×10^{6}	1.07×10^{7}		
Mean Motion	0.000727	0.007215	0.00268	0.00146		
StDev Motion	0.00351	0.01952	0.0107	0.00423		
Shot Lenght	22.3	847.5	101	161		

Figure 1. Basic statistics for low-level video features derived from the ads. The table columns represent the minimum, maximum, mean, and standard deviation for each video feature, respectively. The last two columns show the randomized bar graph of raw consecutive feature values for each ad and their histograms.

Lighting has a nearly uniform distribution around an average value of 0.117. Color Variance has much higher values in the range of 10^6. The values for Mean Motion Activity and Standard Deviation of Motion Activity are low and in the range of 10^{-3}. Average Shot Length of the videos is in the range of 40–80 frames in most cases, with some single-shot videos pushing the average shot duration towards 850 frames. The histogram distribution of Color Variance, Mean Motion Activity, Standard Deviation of Motion Activity and Average Shot Length is uneven, skewed to the right with a strong gap towards the maximum values.

4.1.2. Consumer Engagement Scores

The consumer engagement scores were collected in the observational study using UES-SF. The scores were calculated for the 30 videos and all four UES-SF dimensions (see Section 3). The results for 150 participants are shown in Figure 2.

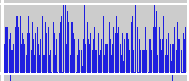
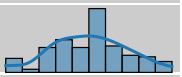
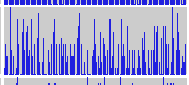
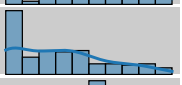
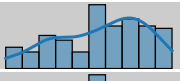
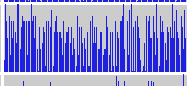
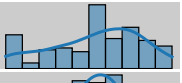
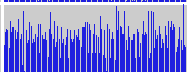
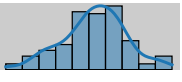
	min	max	mean	std	raw	hist
Focused attention (FA)	1	5	3	0.93		
Perceived usability (PU)	1	4.7	2.2	1		
Aesthetic appeal (AE)	1	5	3.3	1.1		
Reward (RW)	1	5	3.2	1.1		
Overall engagement (UES)	1.1	4.5	2.9	0.65		

Figure 2. Basic statistics for ground truth data collected using UES-SF. The table columns represent the minimum, maximum, mean, and standard deviation for each dimension. The last two columns present a randomized bar graph of raw consecutive values of the engagement dimensions for each video rating and their histograms.

The reliability coefficient hierarchical ω of the overall model is $\omega = 0.81$. Measurement models of all four dimensions of USE-SF are tau-equivalent [33] and the pertained reliability coefficient is Tau-equivalent ρ_T. For comparability, we also report commonly used reliability coefficients Cronbach's α, see Table 3. The reliability of the dimensions Aesthetic Appeal (AE) and Reward (RW) is very high, and the estimated reliability of the dimensions Focused Attention (FA) and Perceived Usability (PU) is moderately high. The obtained values are comparable to the reliability measures reported in the original article presenting UES-SF [5].

Table 3. Reliability coefficients Cronbach α and Tau-equivalent ρ_T for each UES-SF dimension.

UES-SF	Cronbach α	Tau-eq. ρ_T
FA	0.73	0.76
PU	0.70	0.75
AE	0.81	0.83
RW	0.86	0.89

Figure 2 shows the statistics for the UES-SF scores. The scores for the UES-SF dimensions are given on a scale from 1 to 5. The minimum, maximum, mean, and standard deviation are summarized in the first four columns. The randomized bar graphs in the fifth column represent the individual dimension values of engagement bound to each consecutive observation. The last column contains the histograms. The values of Focused Attention (FA) are distributed over the entire score range in a near-normal shape. Perceived Usability (PU) scores tend to express a high number of lowest rates (1), with other values distributed uniformly. The distributions of Aesthetic Appeal (AE) and Reward (RW) tend to dominate in high values between 3 and 5. Overall engagement score (UES) expresses appearance of a normal distribution with a mean value of 3 but of a relatively small deviation.

4.2. Modeling Consumer Ad Engagement

Based on the UES-SF scores and video analysis, we conducted a thorough review of the contributions of low-level video features to each UES-SF dimension. At first, we tried to identify possible linear associations between the video features and UES-SF scores.

4.2.1. Identification of Linear Associations of Video Features and Engagement

The significance of video features in terms of linear association for a given UES-SF dimension was tested using the Kruskal–Wallis ANOVA test (dependent variable normality assumptions of ANOVA were not met, and all Shapiro–Wilks test p-values were < 0.05). Levels for each video feature are determined per feature individually. Results are shown

in Table 4. There were only three positive feature identifications. Regarding the multiple-testing problem [34], the expected number of false positives is $25 \times 0.05 = 1.2$ that is one or possibly two false positives.

Table 4. Reported p-values of Kruskal–Wallis test with randomized effects, where the independent variables are the video features, and the dependent variables are the individual dimensions of the UES-SF scale.

Video Feature	FA	PU	AE	RW	UES
avgLightKeyAd	0.488	0.174	**0.029**	**0.043**	0.192
avgColVarAd	0.218	0.795	**0.032**	**0.017**	0.064
avgMotionMeanAd	0.391	**0.037**	0.689	0.404	0.494
avgMotionStdAd	0.330	0.220	0.270	0.805	0.705
avgShotLenAd	0.221	0.446	0.221	0.396	0.998

4.2.2. Identification of Model Shapes Associating Video Features and Engagement

A series of scatter plots comparing each selected video feature in pairs with the UES-SF dimensions were investigated both visually and statistically to determine the nature of the relationships between them. Visual inspection revealed a potential linear relationship between the low-level visual features and the corresponding dimensions of engagement. For example, the relationship between the video feature Lighting Key and the UES-SF dimension Reward (RW) indicates a linear relationship, as shown in Figure 3a.

Several cases indicate that there may be more complex dependencies in addition to a simple linear relationship. Considering the U-shaped (or possible inverted U-shaped) relationship between a perceived response and its stimulus, which is often found in studies on human behavior [35,36], nonlinear modeling was also considered. At this stage of the study, the goal was to identify the nature of the relationship rather than to define an optimal model. A quadratic model was applied as the simplest nonlinear model capable of capturing non-monotonous relations. An example of a possible U-shaped relationship between Standard deviation of Motion and the UES-SF dimension Perceived Usability (PU) is shown in Figure 3b. Figure 3c represents a U-shaped relationship between the mean value of Color Variance and overall User Engagement Score (UES).

4.3. Contributions of Video Features to UES-SF

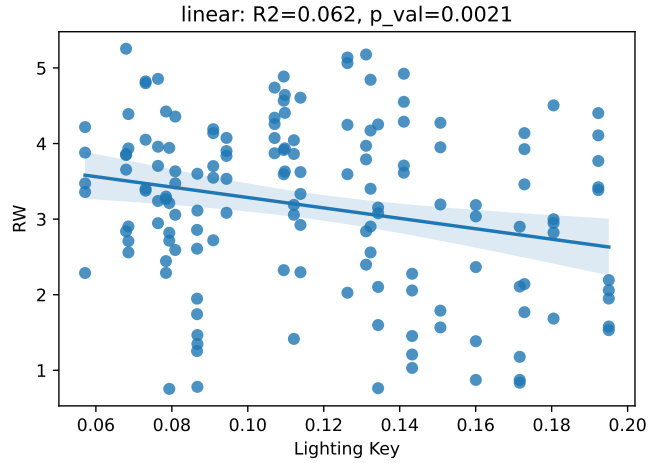
Based on these results, several regression analyses were performed to determine the nature of the relationship between all the possible feature and the consumer engagement rating pairs. For each combination, the p-value of the null hypothesis $H_0 = [R^2 = 0]$ was calculated, where R^2 represented a coefficient of determination of the analyzed model. The assumed risk level was set at $\alpha = 0.05$.

The confidence interval of the fitted curve (confidence curve) using a standard bootstrapping procedure [37] was used to test the stability of the obtained models. The mean curves and confidence curves shown in Figure 3 indicate stable modeling, and there were no cases of instability in all 25 cases of modeling.

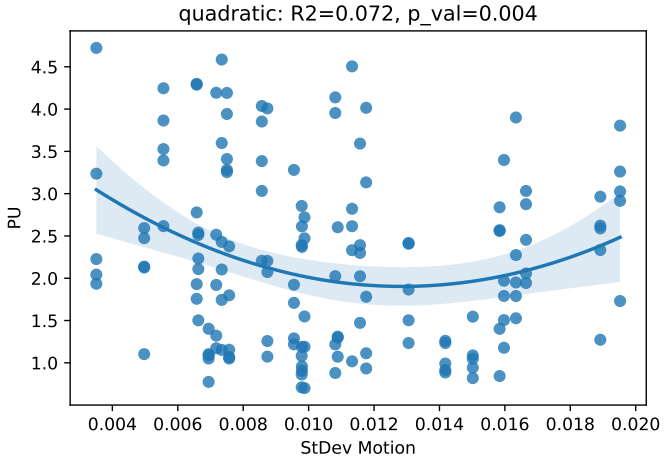
The contributions of individual video features are discussed in the following subsections.

4.3.1. Lighting Key

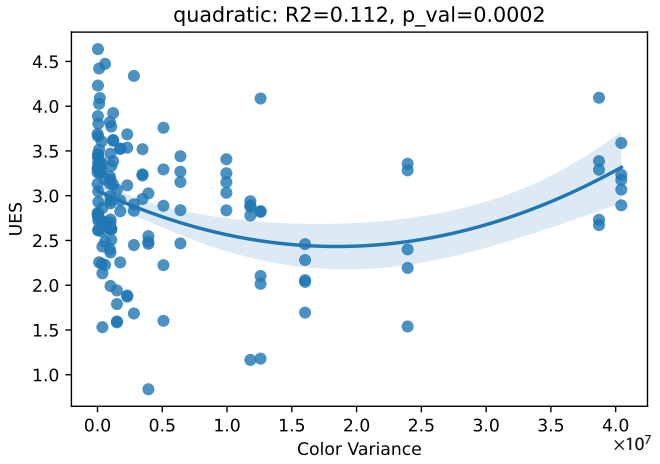
Lighting Key is a video feature that represents the relationship between light and shadow in the video. The regression analyses of this feature for each of the engagement dimensions are shown in Table 5. Significant relations have been found in the case of linear models for the dimensions of Aesthetic Appeal (AE) and Reward (RW) as well as in the case of overall User Engagement Score (UES). Statistical evaluation also confirms the quadratic relation in the dimensions of Aesthetic Appeal (AE) and Reward (RW); although R^2 values are slightly in favour here, upon visual inspection, one can identify little improvements in case quadratic factors are deployed.



(**a**)



(**b**)



(**c**)

Figure 3. Examples of (**a**) scatter plot and linear model plot of a decreasing linear relation between Lighting Key and Reward (RW); (**b**) scatter plot and a quadratic model plot of relation between Color Variance and overall User Engagement Score (UES), indicating a U-shaped relation; (**c**) scatter and quadratic plot of the relation between Standard Deviation of Motion and Perceived Usability (PU) dimension. Please note that a small jitter was added to improve visualisation of overlapped values.

Table 5. Lighting Key in relation to user engagement scale (short form) for the linear and quadratic model.

	Linear		Quadratic	
	R^2	p	R^2	p
FA	0.005	0.375	0.008	0.535
PU	0.006	0.33	0.016	0.311
AE	0.051	<0.01	0.054	0.018
RW	0.062	<0.01	0.067	<0.01
UES	0.037	0.018	0.037	0.062

4.3.2. Color Variance

The video feature Color Variance shows the variance of colors within a video, and its relationship with consumer engagement is shown in Table 6. We did confirm a quadratic relation of the feature in most dimensions, as well as in overall UES score. The strongest confirmed relation was identified for Focused Attention (FA) and Aesthetic Appeal (AE), followed by Reward (RW). There is no relation if this feature with Perceived Usability (PU) could be seen. In all cases, the U-shaped relation has been identified, indicating that either high or low values of Color Variance increase user exposure scores in total and in the reported dimensions of FA, AE and RW.

Table 6. Color Variance relation to user engagement scale (short form) for linear and quadratic model.

	Linear		Quadratic	
	R^2	p	R^2	p
FA	0.0	0.804	0.078	<0.01
PU	0.003	0.479	0.007	0.58
AE	0.0	0.973	0.074	<0.01
RW	0.0	0.998	0.054	0.017
UES	0.001	0.728	0.112	<0.01

4.3.3. Average Shot Length

A characteristic feature of video editing is the average duration of shots, i.e., continuous frame sequences between two cuts. The duration of the shots is related to the dynamics of the video shot or cut. The statistics of this feature are extremely uneven in the case of our video selection because a small number of videos were shot using the one-shot cinema technique and therefore consist of a single long shot, which then results in extreme values of the feature. The regression and scatter plots were therefore not optimally suited for visual inspection. The regression analyses of this feature for each of the engagement dimensions are shown in Table 7. No significant relation between Lighting Key alone and any of the engagement dimensions could be identified from the visualisations, which matches the high reported p-values.

Table 7. Shot Length relation to user engagement scale (short form) for linear and quadratic model.

	Linear		Quadratic	
	R^2	p	R^2	p
FA	0.007	0.302	0.027	0.129
PU	0.006	0.361	0.009	0.511
AE	0.001	0.698	0.034	0.079
RW	0.016	0.118	0.022	0.193
UES	0.005	0.405	0.023	0.182

4.3.4. Motion Mean and Motion Standard Deviation

The degree of motion differs significantly between certain video categories, such as action movies, sports, relaxing movies, and dramas. As explained in Section 2.2.2, we used two different features related to the average motion activity and its variability. The linear and quadratic model fits of engagement dimensions to mean motion computed using

optical flow are summarized in Table 8. Table 9 refers to linear and quadratic models of standard deviation of motion. The values of p for quadratic models of both the mean and standard deviation of the motion were below $\alpha = 0.05$ for the sub-dimension of Perceived Usability (PU). Figure 3c as an example demonstrates a U-shaped relation between the mean value of Motion Activity and Perceived Usability (PU) dimension. The same behaviour could be identified for the case of Motion Standard Deviation. This behaviour indicates optimum values of motion which will lead to the best sub-score of Perceived Usability (PU). This implies that very slow or very fast motion in advertising has a negative impact on this dimension of engagement.

Table 8. Mean motion in relation to user engagement scale (short form) for a linear and quadratic regression model.

	Linear		**Quadratic**	
	R^2	p	R^2	p
FA	0.001	0.749	0.01	0.476
PU	0.004	0.452	0.062	<0.01
AE	0.01	0.225	0.015	0.326
RW	0.01	0.232	0.01	0.464
UES	0.005	0.409	0.012	0.407

Table 9. Standard deviation of motion in relation to user engagement scale (short form) for a linear and quadratic regression model.

	Linear		**Quadratic**	
	R^2	p	R^2	p
FA	0.0	0.859	0.001	0.931
PU	0.017	0.107	0.072	<0.01
AE	0.01	0.214	0.031	0.099
RW	0.007	0.3	0.014	0.349
UES	0.0	0.793	0.001	0.964

4.3.5. Feature Contributions and Model Shapes

Overall, a series of 25 combinations of video features and their respective engagement dimensions including its total score were tested. Each of the five low-level video features was tested for four engagement dimensions and for a total engagement score (see Tables 5–9), using both linear and quadratic models. Based on the statistical significance of the coefficients of determination R^2 (null hypothesis $H_0 = [R^2 = 0]$), the low-level video features were identified as significant in two cases for the linear models and in six cases for the quadratic models.

As indicated in Section 4.2, the polynomial models were restricted to quadratic order. Only the models that were significant in terms of the coefficient of determination were considered, despite the fact that also some other plots exhibit convincing model shapes.

The observed shapes are presented in Table 10. To summarize the relationships between the video features and the predicted engagement dimensions, the relations of the Color Variance video feature towards engagement score subdimensions of Focused Attention (FA), Aesthetic Appeal (AE) and Reward (RW), as well as towards overall User Engagement Score (UES), are U-shaped. The same observations are valid for relations of video features Motion Mean and Motion Standard Deviation with Perceived Usability (PU) and for the video feature Shot Length towards Aesthetic Appeal (AE). We identified no case of a prevailing linear model shape. Typical representations of these shapes are shown in Figure 3. The interpretation of these observations is weak because very little theoretical knowledge is available from the existing studies.

Table 10. Model shapes of four O'Brien UES user engagement scale dimensions estimated from 5 low-level ad video features. The foreseen possible model shapes are labelled with symbols ∪ representing U-shape, ∩ indicating upside-down *U* shape, ↗ showing increasing and ↘ for decreasing behaviour.

	Light Key	Color Var	Shot Lenght	Motion Mean	Motion Std
FA		∪			
PU				∪	∪
AE	↘	∪			
RW	↘	∪			
UES	↘	∪			

In general, the tests yielded low R^2 values, suggesting that the models explained little of the variability in the UES-SF dimensions. These results were expected and can be seen visually in the widely dispersed score values shown in Figure 3 and in the other scatter plots under evaluation. Further discussion of unexplained variability in the models is provided in Section 4.4.

Regarding the multiple-testing problem [34], the expected number of false positives is $25 \times 0.05 = 1.2$, which is one or possibly two false positives.

4.4. Unexplained Variability of Consumer Ad Engagement

Based on the R^2 values, the models presented explain only a small portion of the variability (up to 10%) in the consumer ad engagement scores collected from UES-SF. Figure 3 shows the scatter plots with the six vertical clusters bound to the feature values of each video and plotted on the x-axis.

This unexplained variability is not uncommon, especially in human psychology studies [14]. Moreover, this was expected in the present study because of the nature and complexity of the engagement construct and the lack of experimental data from related work. Consequently, the proposed models based on low-level video features alone cannot achieve a higher value of explained variance (R^2).

To put the reported R^2 values into context, the question is what the contribution is of the evaluated models compared to the maximum achievable value of R_m^2 for a given feature. To determine the maximum achievable R_m^2 for an ideal model based on video features, the ad engagement scores were evaluated using categorical regression based on the ad videos themselves. Categorical data from the six videos were quantified by assigning numerical values to their respective categories with the goal of the model yielding the best possible linear regression fit. The mean value for consumer engagement in ads was used as a separate transformation of the categorical video value for each engagement dimension.

An example of a scatter plot and the corresponding linear regression plot for overall User Engagement Score (UES) is shown in Figure 4. A high level of variability in consumer engagement can be observed for each video.

A summary of maximum achievable R^2 values for the case of categorical regression based on video advertisements themself can be found in Table 11, where values in the data column contain the estimated R_m^2 values of the ideal model. Relatively low values agree with our initial assumption (see Equation (1)) that there are other factors affecting the consumer engagement apart from the content (video) itself.

Table 11. Maximum R^2 values for categorical regression quantified by assigning idealized numerical values for each engagement dimension.

UES-SF	Linear	
	R^2	p
FA	0.264	<0.01
PU	0.402	<0.01
AE	0.44	<0.01
RW	0.506	<0.01
UES	0.365	<0.01

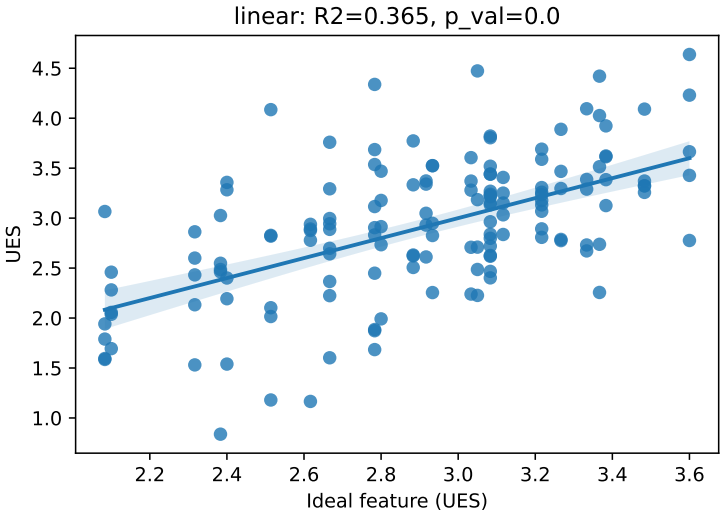


Figure 4. Scatter plot and corresponding linear regression plot for the optimal video feature against the overall User Engagement Score (UES). Note that the x-axis (abscissa) relates to categorical variable of Video ID; the assigned numerical values are equal to corresponding mean engagement scores.

5. Discussion

The article presents a novel approach to model the influence of automatically derived low-level video features on short-term ad engagement. Most existing approaches for modeling consumer ad engagement are based on video metadata and use high-level content features (e.g., genre, director, cast, tags, text ratings, and similar metadata). Such metadata are often scarce and may not capture the impact of visual video features on consumers. The implicit metadata available through the computational analysis of low-level video features might better capture some of the ad's visual effects on consumers. Moreover, in most cases, the existing approaches lack referential ground truth for consumer-perceived engagement.

To this end, UES-SF scores on the four dimensions of consumer engagement were collected to provide the ground truth. Low-level video features were then derived and used to model the UES-SF scores. The results show that several features can be used to estimate consumer engagement, but each explains only a small portion of the variability for the respective UES-SF dimensions.

For most low-level video features studied, we can identify their relationships with the four dimensions of UES-SF. In the existing literature, we have found no linear nor more complex models of user engagement scores based on low-level video descriptors. However, we found some evidence of U-shaped behavior of the models related to human behaviour [35,36]. Indeed, linear models have been shown to explain a statistically significant portion of the variability only in rare cases of our study. Given the tentative U-shaped model options, we opted for quadratic models and identified a satisfactory number of significant model relations.

The upper limit of the explained variability related to the video contribution by any of the models is low for itself, ranging from 25% to 50%, see Table 11. This behaviour, also indicated by relatively small R^2 values, was expected as the factors of personality and context influencing the user engagement were not part of this study. Consumer ad engagement is a complex construct that cannot be captured solely by visual features of an ad, but is influenced by many factors, such as context and consumer perceptions and attitudes, such as personality, brand likability, familiarity, and recall; see Equation (1).

There are several limitations to the presented study. First, as a general limitation, the study addresses only short-term engagement and does not address possible changes or patterns of ad engagement over time. Second, the study participants represent a specific group of consumers within a narrow age range and in a specific location (the United States). This group may have specific consumer behavior and usage patterns, as well as social and/or cultural characteristics that cannot be generalized to other consumer groups or

locations. In addition, only personal computer devices were used in this study, while usage patterns and consumer behavior in general may be different on mobile devices such as tablets and smartphones. Another potential limitation is the choice of UES-SF to collect ground truth data. UES-SF may be too general to provide more detailed insights into consumer engagement, such as brand awareness and recall, which are important factors in advertising. This potential problem needs to be explored in the future, as the current state of the art does not provide comparable psychometric instruments for measuring consumer engagement in in-video ads.

6. Conclusions

The presented approach is robust and scalable to other domains where video features could be used to model engagement. It is also novel compared to the existing literature. While most existing approaches model consumer engagement based on some descriptive metadata (such as views, like/dislike ratio, time spent watching, sentiment, etc.), the presented approach automatically models consumer engagement from the derived low-level video features of an ad. Moreover, the proposed estimation of ad engagement is based on the ground truth of perceived engagement measured by the established psychometric instrument, the UES-SF. To this end, the presented estimates of engagement and video feature importances are perceptually grounded.

The presented observational study is part of a broader and more complex investigation of consumer behavior related to in-video advertising. Overall, the results presented are encouraging. However, further studies with a larger number of materials and participants will be conducted to further evaluate the presented findings.

In the future, we will also investigate how user-related factors with varying (e.g., mood, fatigue) and non-varying factors (personality) can improve ad engagement estimates. To this end, our future work will aim to take a more holistic approach by incorporating context and content metadata to build a more complete model of user engagement.

Author Contributions: Conceptualization, E.A.O. and A.K.; Data curation, E.A.O. and U.B.; Formal analysis, G.S. and A.K.; Funding acquisition, A.K.; Investigation, G.S. and U.B.; Methodology, G.S., A.K. and U.B.; Project administration, A.K.; Resources, E.A.O. and U.B.; Software, G.S., A.K. and U.B.; Supervision, A.K.; Validation, G.S., A.K. and U.B.; Visualization, U.B.; Writing—original draft, U.B.; Writing—review and editing, E.A.O., G.S. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the projects MMEE—Multimedia Exposure Estimation and P2-0246 ICT4QoL—Information and Communications Technologies for Quality of Life.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board. All procedures performed were in accordance with the ethical standards of the institutional research committee.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: A collection of scripts and user engagement data used in this research is available at https://doi.org/10.5281/zenodo.7341316, accessed on 20 November 2022. For copyright issues, the access of video advertisements is available only in the form of YouTube links. Readers may refer to description files on GitHub for more information.

Acknowledgments: The authors would like to thank The Nielsen Company (US), LLC, https://www.nielsen.com/, accessed on 20 November 2022, for marketing expert cooperation and valuable opinions regarding the target audience and the different contexts of potential consumer engagement.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UES	User Engagement Scale
UES-SF	User Engagement Scale-Short Form [5]
FA	Focused Attention
AE	Aesthetic Appeal
PU	Perceived Usability
RW	Reward

References

1. Danaher, P.J. Advertising Effectiveness and Media Exposure. In *Handbook of Marketing Decision Models*; Wierenga, B., van der Lans, R., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 463–481. [CrossRef]
2. Calder, B.J.; Isaac, M.S.; Malthouse, E.C. How to Capture Consumer Experiences: A Context-Specific Approach To Measuring Engagement. *J. Advert. Res.* **2016**, *56*, 39–52. [CrossRef]
3. Dessart, L.; Veloutsou, C.; Morgan-Thomas, A. Capturing consumer engagement: Duality, dimensionality and measurement. *J. Mark. Manag.* **2016**, *32*, 399–426. [CrossRef]
4. Araujo, T.; Copulsky, J.R.; Hayes, J.L.; Kim, S.J.; Srivastava, J. From Purchasing Exposure to Fostering Engagement: Brand–Consumer Experiences in the Emerging Computational Advertising Landscape. *J. Advert.* **2020**, *49*, 428–445. [CrossRef]
5. O'Brien, H.L.; Cairns, P.; Hall, M. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *Int. J. -Hum.-Comput. Stud.* **2018**, *112*, 28–39. [CrossRef]
6. Che, X.; Ip, B.; Lin, L. A Survey of Current YouTube Video Characteristics. *IEEE Multimed.* **2015**, *22*, 56–63. [CrossRef]
7. Nijholt, A.; Vinciarelli, A. Measuring engagement: Affective and social cues in interactive media. In Proceedings of the 8th International Conference on Methods and Techniques in Behavioral Research, Measuring Behavior, Utrecht, The Netherlands, 28–31 August 2012; Volume 63, pp. 72–74.
8. Hollebeek, L.; Glynn, M.; Brodie, R. Consumer Brand Engagement in Social Media: Conceptualization, Scale Development and Validation. *J. Interact. Mark.* **2014**, *28*, 149–165. [CrossRef]
9. Shen, W.; Bai, H.; Ball, L.J.; Yuan, Y.; Wang, M. What makes creative advertisements memorable? The role of insight. *Psychol. Res.* **2020**, *85*, 2538–2552. [CrossRef]
10. Niederdeppe, J. Meeting the Challenge of Measuring Communication Exposure in the Digital Age. *Commun. Methods Meas.* **2016**, *10*, 170–172. [CrossRef]
11. De Vreese, C.H.; Neijens, P. Measuring Media Exposure in a Changing Communications Environment. *Commun. Methods Meas.* **2016**, *10*, 69–80. [CrossRef]
12. Gambetti, R.C.; Graffigna, G.; Biraghi, S. The Grounded Theory Approach to Consumer-brand Engagement: The Practitioner's Standpoint. *Int. J. Mark. Res.* **2012**, *54*, 659–687. [CrossRef]
13. Anderson, A.; Hsiao, T.; Metsis, V. Classification of Emotional Arousal During Multimedia Exposure. In Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, Rhodes, Greece, 21–23 June 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 181–184. . [CrossRef]
14. Hoyt, W.T. Rater bias in psychological research: When is it a problem and what can we do about it? *Psychol. Methods* **2000**, *5*, 64. [CrossRef] [PubMed]
15. Chaturvedi, I.; Thapa, K.; Cavallari, S.; Cambria, E.; Welsch, R.E. Predicting video engagement using heterogeneous DeepWalk. *Neurocomputing* **2021**, *465*, 228–237. [CrossRef]
16. Bulathwela, S.; Pérez-Ortiz, M.; Lipani, A.; Yilmaz, E.; Shawe-Taylor, J. Predicting engagement in video lectures. *arXiv* **2020**, arXiv:2006.00592.
17. Stappen, L.; Baird, A.; Lienhart, M.; Bätz, A.; Schuller, B. An Estimation of Online Video User Engagement From Features of Time- and Value-Continuous, Dimensional Emotions. *Front. Comput. Sci.* **2022**, *4*, 773154. [CrossRef]
18. Wu, S.; Rizoiu, M.A.; Xie, L. Beyond Views: Measuring and Predicting Engagement in Online Videos. In Proceedings of the International AAAI Conference on Web and Social Media, Palo Alto, CA, USA, 25–28 June 2018; Volume 12.
19. Lops, P.; de Gemmis, M.; Semeraro, G. Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*; Springer: New York, NY, USA, 2010; pp. 73–105. [CrossRef]
20. *ISO/IEC TR 15938*; ISO Central Secretary. Standard ISO/IEC TR 15938. International Organization for Standardization: Geneva, Switzerland, 2002.
21. Martinez, J.M.; Koenen, R.; Pereira, F. MPEG-7: The generic multimedia content description standard, part 1. *IEEE Multimed.* **2002**, *9*, 78–87. [CrossRef]
22. Baştan, M.; Çam, H.; Güdükbay, U.; Özgür Ulusoy. BilVideo-7: An MPEG-7-Compatible Video Indexing and Retrieval System. *IEEE Multimed.* **2009**, *17*, 62–73. [CrossRef]
23. Eidenberger, H. Statistical analysis of content-based MPEG-7 descriptors for image retrieval. *Multimed. Syst.* **2004**, *10*, 84–97. [CrossRef]
24. Zettl, H. Essentials of Applied Media Aesthetics. In *Media Computing*; Springer: New York, NY, USA, 2002; pp. 11–38. [CrossRef]

25. Ricci, F.; Rokach, L.; Shapira, B. Introduction to Recommender Systems Handbook. In *Recommender Systems Handbook*; Springer: New York, NY, USA, 2010; pp. 1–35. [CrossRef]
26. Deldjoo, Y.; Elahi, M.; Cremonesi, P.; Garzotto, F.; Piazzolla, P.; Quadrana, M. Content-Based Video Recommendation System Based on Stylistic Visual Features. *J. Data Semant.* **2016**, *5*, 99–113. [CrossRef]
27. Deldjoo, Y.; Elahi, M.; Quadrana, M.; Cremonesi, P. Using visual features based on MPEG-7 and deep learning for movie recommendation. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 207–219. [CrossRef]
28. Rasheed, Z.; Sheikh, Y.; Shah, M. On the use of computable features for film classification. *IEEE Trans. Circuits Syst. Video Technol.* **2005**, *15*, 52–64. [CrossRef]
29. Kobla, V.; Doermann, D.; Faloutsos, C. Video Trails: Representing and Visualizing Structure in Video Sequences. In Proceedings of the Fifth ACM International Conference on Multimedia—MULTIMEDIA '97, Seattle, WA, USA, 9–13 November 1997. [CrossRef]
30. Horn, B.K.; Schunck, B.G. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203. [CrossRef]
31. Vasconcelos, N.; Lippman, A. Statistical models of video structure for content analysis and characterization. *IEEE Trans. Image Process.* **2000**, *9*, 3–19. [CrossRef] [PubMed]
32. Home-PySceneDetect. Available online: https://pyscenedetect.readthedocs.io/en/latest/ (accessed on 18 March 2022).
33. Cho, E. Making Reliability Reliable: A Systematic Approach to Reliability Coefficients. *Organ. Res. Methods* **2016**, *19*, 651–682. [CrossRef]
34. Miller, R.G. *Simultaneous Statistical Inference*; Springer: New York, NY, USA; Heidelberger/Berlin, Germany, 1981.
35. Northoff, G.; Tumati, S. Average is good, extremes are bad—Nonlinear inverted U-shaped relationship between neural mechanisms and functionality of mental features. *Neurosci. Biobehav. Rev.* **2019**, *104*, 11–25. [CrossRef]
36. Van Steenbergen, H.; Band, G.P.H.; Hommel, B. Does conflict help or hurt cognitive control? Initial evidence for an inverted U-shape relationship between perceived task difficulty and conflict adaptation. *Front. Psychol.* **2015**, *6*, 00974. [CrossRef]
37. Tibshirani, R.J.; Efron, B. An introduction to the bootstrap. *Monogr. Stat. Appl. Probab.* **1993**, *57*, 1–436.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.