Improving the Performance of Multiscene Marketing Video Content through its Dynamics Adjustments

Kacper Fornalczyk, Kamil Bortko, Aneta Disterheft, Jarosław Jankowski
0000-0003-3752-3473
0000-0002-3658-3039
West Pomeranian University of Technology
Faculty of Computer Science and Information Technology
Żołnierska 49
71-210 Szczecin, Poland
Email: kacper_fornalczyk@wp.pl, anetadisterheft@gmail.com, {kbortko, jjankowski}@zut.edu.pl

Abstract—The use of online video content plays a vital role in marketing strategies and is a significant component of internet usage. The challenge lies in evaluating the impact of video content on user engagement and finding ways to enhance its performance without employing techniques that overwhelm users or prompt ad avoidance behavior. This study investigates the correlation between video dynamics metrics and eye-tracking patterns to determine if user engagement, as indicated by fixations, is influenced by these metrics. The findings demonstrate that dynamic metrics can accurately predict eye-tracking patterns for brief videos and can be applied to measure both inter and intra-scene dynamics in multiscene videos.

I. Introduction

ONLINE video content is a popular medium that comprises a significant share of internet usage, with its consumption expected to rise to 82% of all internet traffic by 2022, up from 75% in 2017, according to Cisco's 2018 report [6]. Video content is widely used for marketing purposes, in the form of in-stream ads or integrated with editorial content on social platforms, games, or portals[17]. Content creators often use techniques that increase user engagement through emotional content, visual effects, and high dynamics, but these techniques can also increase cognitive load and distract users from their main goals within websites[14]. This decreased user experience may lead to users skipping advertising content, particularly when it fails to catch their attention at first glance. Hence, content producers face the challenge of creating video ads that are less likely to be skipped by consumers, which can be achieved by lowering intrusiveness and the dynamics of the video content. This paper investigates how the dynamics of video, as represented by dedicated metrics, relate to eye-tracking patterns, and whether the dynamics of video can predict user engagement, as represented by fixations. The secondary goal is to examine the impact of intra-scene differences on user attention within multi-scene videos. The primary aim is to explore methods of building videos with low

dynamics and low cognitive load while maintaining enough differences between them to sustain user attention.

II. Literature review

Although there is already some research on television advertising, there is still an opportunity to delve deeper into the area of online video ads, as this format is more captivating and hence more widely used than static display or text ads [19]. Advertisers and advertising area providers need to know when video ads start to become too distracting, which can result in the use of ad blockers. This situation underscores the need to search for factors that affect video ad performance [20], especially those that can be used to attract users' attention and keep them engaged [5] [11]. High cognitive load may result from different ad characteristics, such as high video dynamics or the use of intense colors or sounds that are perceived as disturbing[15]. The following research focuses on measuring video ad dynamics as a factor affecting performance. An algorithm that automatically extracts several features, including video-level visual variance, scene-to-scene visual variance, and average scene cut frequency, was developed and published in [13] and further elaborated in [19] [2]. As the algorithm is publicly available, it was used in the following study.

One concept that plays a significant role in this notion is the level of involvement shown by consumers. Numerous writers have consistently reported that viewers who are more engaged in video content are less likely to skip it [10][21]. Other studies in this field have highlighted the correlation between ad avoidance and cognitive factors that are closely linked to engagement. For instance, in [1] [4], it was discovered that avoiding ad content may be connected to high and low arousal levels elicited by the content. A comparable finding was outlined in [18], which revealed that unstimulating, uninteresting content increases the desire to skip ads[9]. The primary factor influencing consumers' tendency to avoid ads is the engagement level of the ad content. Additionally, the length

Thematic track: Information Systems Management

of the video ad is another critical factor that affects ad avoidance [18]. Numerous studies have discussed the relationship between ad length and the rate of ad avoidance, identifying a correlation between increased skipping behaviour and longer content [18] [12]. Other research has suggested that longer ads result in more disruption for goal-oriented search [9][3]. Recent studies have revealed that consumers' acceptance of longer videos has decreased, and nowadays, most people only accept very short video ads, such as fifteen or six seconds in length [18]. As a result, the current trend is to produce short video marketing content that better targets consumers' attention spans[7][8]. This trend is advantageous for ad providers because they can increase the rate of presenting content to consumers without increasing the total costs of a campaign. However, some studies have shown that longer ads may be more effective in enhancing brand recognition [12]. Ads that are shortened to fifteen seconds can achieve similar results in terms of awareness and brand recall as thirty-second spots[16].

In this current study, various factors were examined to evaluate the performance of videos. We aimed to investigate how the video's dynamics and differences between scenes can affect eye-tracking patterns and how user engagement relates to video dynamics, both at the level of single videos and intra-scene differences.

Conceptual framework

The conceptual drawing Figure 1 shows the stage 1 in which you can see the prepared video base divided into the length of the films and their dynamics. After preparation of the base, the test was carried out in laboratory conditions, with the intention of obtaining an increased number of fixations along with increasing dynamics.

Stage 2 presents the users' gaze pattern thanks to the eyetracker examination of interstage dynamics. You can see here an increase in dynamics in relation to the increasing number of fixations for short scenes versus long scenes, where such a relationship does not exist.

III. Algorithm description

To study video dynamics we implemented algorithm parented in Xi Li, Mengze Shi and Xin (Shane) Wang [13] and we extended it towards measuring inter and intra-scene dynamics. The input of the algorithm apart from file is the list of numbers being the numbers of consecutive frames that constitute the beginning of each new scene. Technically, a scene is a group shots which are successively taken together at a single location. A shot is a basic narrative element of the video which is composed of a number of frames that are presented from a continuous viewpoint. Automatically dividing a video into its shots is called the shot boundary detection problem in which the basic idea is identifying consecutive frames that form a transition from one shot to another. Currently, there are more or less effective solutions to this problem that can be used to obtain the above-mentioned list.

Additionally, one of the two possible parameters should be specified at the input of the algorithm. One of them takes numerical values ranging from [1, 100], and this value determines what percentage of all frames from each scene should be included in the calculation. For example, if the scene has 200 frames and the parameter value is 20, then 0.2 * 200 = 40 frames, possibly equally spaced from each other, will be extracted from the scene. The second parameter takes values greater than 0 and is an alternative to the previously described parameter. Its value determines the length of the time interval, which is the frequency with which the frame for the analysis will be extracted from the scene. If the scene is 10 seconds long and the parameter is set to 0.5 seconds, then 10 / 0.5 = 20 frames will be set in the scene.

In the first step, the algorithm loads the movie and run through its frames. Each frame is stored in the algorithm's memory as a three-dimensional matrix with dimensions equal to the resolution of the video. Each of the three layers of this three-dimensional matrix contains values that define one of the components of the RBG space for each pixel in the frame. Knowing the numbers of the first frames of all detected scenes, algorithm extracts the appropriate number of frames from each scene with a given interval or percentage, creating a new list for each scene containing the frames extracted for it.

The very measurement of dynamics of a video message is based on the measure provided by Xi Li, Mengze Shi and Xin (Shane) Wang [13]. In their work, the authors present the measure they named "visual variation" which is a normalized measure of the changes in visual information in a video. Determining the visual variation for two frames is carried out in the following few steps.

First, the frames are reduced from RGB to grayscale by averaging the color components of each pixel. The next step is to normalize the values from the range [0, 255] to the range [0, 1]. Authors of the measure mention that normalization serves the purpose of compensation for possible exposure difference. Further, the distance between the individual pixels of the two frames is calculated, where the distance is defined as the Manhattan norm

$$d(x'_i, y'_i) = |x'_i - y'_i| \tag{1}$$

After calculating the matrix with dimensions equal to the resolution of the compared frames, where each position in the matrix is the distance between individual pixels at the same position, the algorithm proceeds to the last step. Here our implementation of the measure differs from the one proposed by its authors. In the original implementation of this measure, in this step of the algorithm, all the determined absolute distances between each pair of pixels should be summed up. In our version, we chose to calculate the mean over the absolute distance of every pair of pixels. This change was aimed at obtaining the visual variation result in the range [0, 1]. This averaged value represents the size of visual variation between two frames.

In the next step determination of the internal dynamics of scenes takes place. This step consists in determining for each list containing extracted frames from individual scenes the average visual variation occurring between consecutive frames
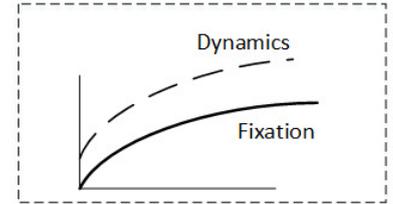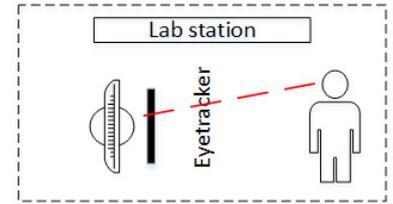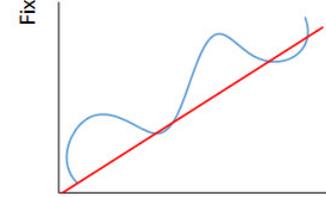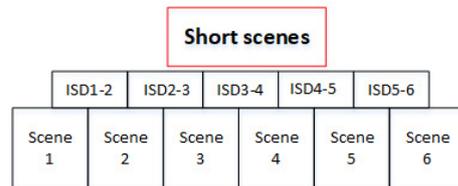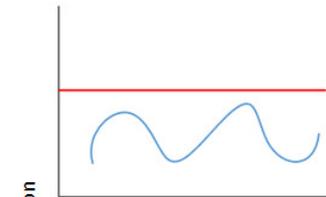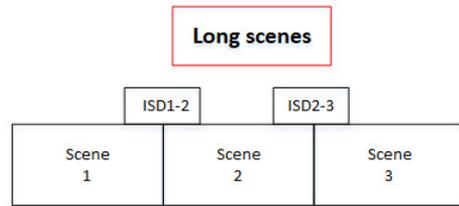
Fig. 1. Conceptual framework for study

in the list. For example, if only 3 frames have been extracted from the scene, then we calculate the value of visual variation between frames no. 1 and no. 2, and between frames no. 2 and no. 3. Then both values are averaged. The process could be written as follows

$$\text{VVL} = \frac{1}{n} \sum_{i=1}^{n} d(F_i, F_{i+1}) \qquad (2)$$

The equation of Visual Variation Level (VVL) where n is the number of frames, F_i is the frame number in the list and d is a function that returns the visual variation between two frames.

A. Determining the external dynamics between the scenes

External dynamics is the working name for the visual variation determined between successive scenes. The calculation of

this measure takes place when the algorithm extracts appropriate frames from individual scenes into new lists. The algorithm knows the order of the lists, which corresponds to the order in which the scenes appear in the entire video transmission. Thanks to this, it can determine the visual variation between individual scenes in the same way as it was presented for a series of frames extracted from one scene.

The first step in determining external dynamics is to average the colors of all frames within each list. The averaged values should be rounded off as the values of the three-dimensional matrix should be integers in the range [0, 255]. This process can be represented by the following formula:

$$Af = [\frac{1}{n} \sum_{i=1}^{n} F_i] \qquad (3)$$

Where F_i - a frame in the form of a three-dimensional matrix, n - the number of frames in a leaf, Af - the resulting averaged three-dimensional matrix.

The above step is performed for each list corresponding to a single scene. This way the algorithm comes to a point where it has an averaged three-dimensional matrix / frame for each scene. Now these averaged frames / matrices can be treated as ordinary frames for which the visual variation within one scene was calculated in the previous section. Here, however, it is done each time only between two averaged frames corresponding to two consecutive scenes (for example, the algorithm calculates the visual variation between the averaged frame from the first scene - s1 - and from the second scene - s2, then between s2 and s3, then s3 a s4... sn-1 a sn). Here, as before, each of the averaged frames is previously reduced to grayscale, normalized and the distances between individual pixels at corresponding positions are calculated, and finally the average is determined from these distances.

The algorithm returns the determined internal dynamics for each scene and external dynamics between successive scenes, as well as additional information about the total number of frames in the examined video, the length of the recording, the number of frames per second, the number of scenes, and the values of the set parameters.

EXPERIMENT AND RESULTS

The experiment involved 30 people involved. The research group consists of 14 women and 16 men aged 20 to 40 years. The experiment used a 27-inch Dell monitor and the Tobii Pro X3 eyetracker with a sampling frequency of 120 Hz. A special stand with a tripod was prepared for the experiment, which made it possible to keep the head of each participant in a stationary position. The participant sat in front of the monitor at a distance of about 54 cm. Calibration was performed before each test.

The whole experiment was as follows: each person performed the task of clicking on points in accordance with the concept of Fitts' law. This task was a kind of a break between the screening of individual films prepared by us. These films were divided into 1, 3 or 6 scenes. Each of the films lasted 15 seconds. The important thing is the variety of internal and external dynamics of films. The internal dynamics have been divided into three levels: low, mid and high. The dynamics measures were determined thanks to the above-described algorithm. So the films were prepared in such a way that, depending on the number of scenes, each of them had the same dynamics.

The films have been divided into 21 combinations, as shown in the Table I The films of 3 and 6 scenes had dynamics measures individually for each of them. The division with respect to the internal dynamics was made additionally to the external dynamics between the scenes, hence so many combinations. Here we see diversity in terms of individual films and scenes. At first glance, the measures and the average number of fixations increase in line with the increase in the dynamics measure for individual scenes in specific movies.
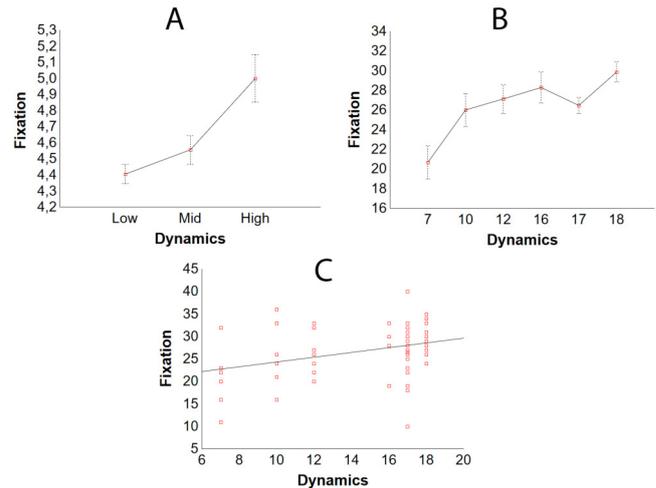
Fig. 2. Fig. A shows a curve prepared with the use of ANOVA statistics showing the trend of the increase in dynamics against the number of fixations for individual scenes. Fig. B shows the total dynamics of scenes showing the increase in dynamics in relation to the number of fixations for entire movies with 6 scenes. Fig. C shows a scatterplot for movie scenes with 6 scenes.
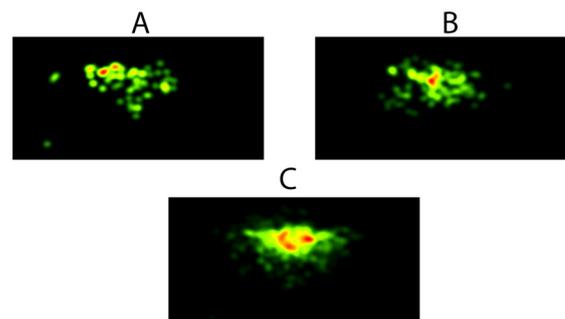

Fig. 3. Heatmaps with visible differences between the various dynamics of the film. Heatmaps A, B, C show three dynamics of movies, respectively: low, mid and high.

Movies and scenes with relatively the highest internal and external dynamics have the best average number of fixations. Here it should be mentioned that external dynamics can only be made for movies that have been divided into a plural number of scenes, ie more than one. Therefore, movies with one default scene were not taken into account in determining the external dynamics measures. As you can see in the Table I, the measure columns for external dynamics are defined for transitions between specific scenes. This allowed to define the dynamics just between them. Figure 2 (A) shows the influence of one intergroup factor, which is the number of fixations per scene, on the dependent variable which is the appropriate measure of dynamics. We can notice a clear upward trend in the number of fixations in relation to the increase in the internal dynamics of films. The number of fixations increases, respectively, from the average to the value of 4.4 for the low dynamics, 4.5 for the mid dynamics, up to the level of 5.0. Figure 2 (B) It shows the total dynamics of scenes for

TABLE I
MEASURES OF DYNAMICS OF EACH SCENE OF INDIVIDUAL FILMS IN EXTERNAL AND INTERNAL DYNAMICS.

film	internal dynamics						external dynamics				
	s1	s2	s3	s4	s5	s6	s1-s2	s2-s3	s3-s4	s4-s5	s5-s6
1_l	0,031										
1_m	0,096										
1_h	0,155										
3_l_l	0,003	0,005	0,005				0,023	0,018			
3_l_m	0,050	0,048	0,057				0,106	0,130			
3_l_h	0,004	0,002	0,006				0,223	0,208			
3_m_l	0,061	0,067	0,060				0,068	0,121			
3_m_m	0,099	0,085	0,064				0,115	0,103			
3_m_h	0,107	0,059	0,088				0,175	0,412			
3_h_l	0,116	0,110	0,119				0,089	0,077			
3_h_m	0,117	0,151	0,136				0,102	0,107			
3_h_h	0,164	0,150	0,147				0,192	0,290			
6_l_l	0,016	0,008	0,017	0,014	0,009	0,021	0,049	0,047	0,039	0,040	0,052
6_l_m	0,046	0,054	0,069	0,037	0,063	0,033	0,174	0,146	0,169	0,131	0,138
6_l_h	0,027	0,013	0,046	0,003	0,028	0,005	0,234	0,233	0,223	0,230	0,230
6_m_l	0,088	0,060	0,078	0,076	0,091	0,078	0,093	0,100	0,091	0,067	0,078
6_m_m	0,074	0,074	0,083	0,079	0,078	0,068	0,163	0,184	0,185	0,137	0,106
6_m_h	0,076	0,077	0,059	0,092	0,079	0,039	0,391	0,376	0,333	0,374	0,189
6_h_l	0,112	0,162	0,130	0,095	0,063	0,089	0,124	0,080	0,079	0,095	0,122
6_h_m	0,105	0,096	0,159	0,144	0,147	0,119	0,171	0,155	0,100	0,091	0,101
6_h_h	0,117	0,110	0,143	0,105	0,114	0,110	0,172	0,171	0,170	0,223	0,184

TABLE II
MEASURES THE AVERAGE NUMBER OF FIXATIONS ON THE SCENE OF A
GIVEN MOVIE.

film	fixations per scene					
	s1	s2	s3	s4	s5	s6
1_l	4,278					
1_m	4,313					
1_h	3,591					
3_l_l	3,286	3,333	3,286			
3_l_m	3,000	3,167	3,000			
3_l_h	3,148	3,111	3,222			
3_m_l	2,714	2,571	3,095			
3_m_m	2,556	3,333	3,000			
3_m_h	2,267	2,800	2,867			
3_h_l	2,833	2,667	2,708			
3_h_m	2,833	3,222	3,778			
3_h_h	2,909	3,152	3,152			
6_l_l	4,667	5,167	5,000	5,667	5,500	5,167
6_l_m	4,000	4,429	4,286	5,000	4,714	5,143
6_l_h	4,333	4,167	3,833	4,333	4,667	4,667
6_m_l	3,500	4,375	3,625	4,125	4,375	4,625
6_m_m	3,111	4,444	4,444	4,333	4,000	5,000
6_m_h	3,571	3,857	4,714	4,571	4,571	4,286
6_h_l	3,900	4,400	4,700	4,300	4,900	5,000
6_h_m	3,800	4,700	4,600	5,300	5,100	5,100
6_h_h	4,143	4,857	4,857	4,429	4,857	4,714

individual films. Here we see an increase in the number of fixations in line with the increase in dynamics in the movies. The trend is clearly increasing, i.e. the number of fixations increases proportionally to the total dynamics. We can see here the ranges for the weakest dynamics on the 7th level of the total number of almost 21 fixations, up to the range of 26 to 30 for the rest of the dynamics. Figure 2 (C) scatter regression plot shows the data in particular dynamics. The presented trend is clearly increasing. Figure 3 shows heatmaps with different dynamics variants. Heatmap A shows the dynamics of Low, the heatmap B shows the dynamics of mid, and the heatmap C the dynamics of high. You can see dense clusters for the high dynamics compared to the other two dynamics.

In this case, the analysis was based on the analysis of internal dynamics for individual films with high, medium and low dynamics. Also in this case, we see the differences between the individual films in each group of internal dynamics broken down by external dynamics.

Using Anova analysis, we see the significance for each dynamics in relation to the number of fixations for individual videos. The significance at the level of $p = 0.017$ indicates a significant dependence of the dynamics measures in relation to each film with the appropriate amount of fixations.

In order to standardize the measures, a division into three groups of dynamics was made using the cluster analysis, which made it possible to reliably and efficiently organize and standardize the groups of individual measures.

Mann - Whitney U statistical analysis, we can see that the intergroup comparison shows a statistical significance below $p < 0.05$ in a few cases. Performing an intergroup comparison here for 6 scene films only showed the significance of w between each group of dynamics. You can see strong differences between the low, mid and high dynamics.

Anova's analysis showed that the summary analysis of scenes for each type of movie shows a significance of $p < 0.05$, which is $p = 0.044$. This shows the strong influence of the amount of fixation on a given scene in the movie.

IV. CONCLUSIONS

Effective video content requires the integration of various elements that increase user engagement, such as emotional appeal, dynamic visuals, and attention-catching techniques. However, the extensive use of video content for marketing purposes has led to avoidance behaviors, such as video ad

skipping or blocking. The acceptable length of videos for users has reduced to just a few seconds. Therefore, it is crucial to develop methods that allow the creation of effective content without sacrificing user experience. In our proposed approach, we demonstrated how metrics of video dynamics are correlated with eye-tracking patterns and can be used to create video content using scenes with different dynamics. By using a modified algorithm that determines the dynamics of individual films and scenes, we correlated these dynamics with the number of fixations while watching them. Our experiment's results showed that dynamic metrics as a predictor of eye-tracking patterns are effective for short videos and can be used for multi-scene films to measure dynamics between and within scenes. The statistics clearly showed an increase in the number of fixations in relation to the increase in dynamics, indicating a directly proportional relationship.

Moving forward, to address the complex challenge of combining these factors in a hybrid approach, in future we propose a framework that integrates both qualitative scene analysis and quantitative visual intensity measurements.

The hybrid approach will first involve a comprehensive scene analysis, where various elements such as objects, shapes, colors, and spatial relationships will be identified and categorized. This qualitative understanding of the scene will provide valuable context for the subsequent analysis. Nonetheless, we believe that this hybrid approach has the potential to enhance our understanding of visual perception and contribute to various fields, such as computer vision, human-computer interaction, and visual design. Through continued research and refinement, the proposed framework could open new avenues for investigating human visual perception and its applications.

Based on these findings, we recommend that advertisers create video content from short films to maximize user absorption. In the future, research will focus on identifying not only the characteristics of scenes based on color differences but also the characteristics of objects within scenes. This will allow for the evaluation of differences based on scene elements, not just visual intensities.

V. Acknowledgements

References

[1] Belanche, D., Flavián, C., Pérez-Rueda, A.: Understanding interactive online advertising: Congruence and product involvement in highly and lowly arousing, skippable video ads. Journal of Interactive Marketing **37**, 75–88 (2017)

[2] Belanche, D., Flavián, C., Pérez-Rueda, A.: User adaptation to interactive advertising formats: The effect of previous exposure, habit and time urgency on ad skipping behaviors. Telematics and Informatics **34**(7), 961–972 (2017)

[3] Campbell, C., Mattison Thompson, F., Grimm, P.E., Robson, K.: Understanding why consumers don't skip pre-roll video ads. Journal of Advertising **46**(3), 411–423 (2017)

[4] Carretié, L., Hinojosa, J.A., Mercado, F.: Cerebral patterns of attentional habituation to emotional visual stimuli. Psychophysiology **40**(3), 381–388 (2003)

[5] Chaturvedi, I., Thapa, K., Cavallari, S., Cambria, E., Welsch, R.E.: Predicting video engagement using heterogeneous deepwalk. Neurocomputing **465**, 228–237 (2021)

[6] Cisco, V.: Cisco visual networking index: Forecast and trends, 2017–2022. White paper **1**(1) (2018)

[7] Codispoti, M., De Cesarei, A., Biondi, S., Ferrari, V.: The fate of unattended stimuli and emotional habituation: Behavioral interference and cortical changes. Cognitive, Affective, & Behavioral Neuroscience **16**(6), 1063–1073 (2016)

[8] Ferrari, V., Mastria, S., Codispoti, M.: The interplay between attention and long-term memory in affective habituation. Psychophysiology **57**(6), e13572 (2020)

[9] Hegner, S.M., Kusse, D.C., Pruyn, A.T.: Watch it! the influence of forced pre-roll video ads on consumer perceptions. In: Advances in Advertising Research (Vol. VI), pp. 63–73. Springer (2016)

[10] Jeon, Y.A.: Skip or not to skip: Impact of empathy and ad length on viewers' ad-skipping behaviors on the internet. In: International Conference on Human-Computer Interaction. pp. 261–265. Springer (2018)

[11] Jeon, Y.A., Son, H., Chung, A.D., Drumwright, M.E.: 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 (2019)

[12] Li, H., Lo, H.Y.: Do you recognize its brand? the effectiveness of online in-stream video advertisements. Journal of advertising **44**(3), 208–218 (2015)

[13] Li, X., Shi, M., Wang, X.S.: Video mining: Measuring visual information using automatic methods. International Journal of Research in Marketing **36**(2), 216–231 (2019)

[14] Mukherjee, A., Dubé, L.: Mixing emotions: The use of humor in fear advertising. Journal of Consumer Behaviour **11**(2), 147–161 (2012)

[15] Myrick, J.G., Oliver, M.B.: Laughing and crying: Mixed emotions, compassion, and the effectiveness of a youtube psa about skin cancer. Health communication **30**(8), 820–829 (2015)

[16] Pace-Schott, E.F., Shepherd, E., Spencer, R.M., Marcello, M., Tucker, M., Propper, R.E., Stickgold, R.: Napping promotes inter-session habituation to emotional stimuli. Neurobiology of learning and memory **95**(1), 24–36 (2011)

[17] Quach, S., Septianto, F., Thaichon, P., Chiew, T.M.: Mixed emotional appeal enhances positive word-of-mouth: The moderating role of narrative person. Journal of Retailing and Consumer Services **62**, 102618 (2021)

[18] Raditya, D., Gunadi, W., Setiono, D., Rawung, J.: The effect of ad content and ad length on consumer response towards online video advertisement. The Winners **21**(2), 119–128 (2020)

[19] Schwenzow, J., Hartmann, J., Schikowsky, A., Heitmann, M.: Understanding videos at scale: How to extract insights for business research. Journal of Business Research **123**, 367–379 (2021)

[20] Semerádová, T., Weinlich, P.: The (in) effectiveness of in-stream video ads: Comparison of facebook and youtube. In: Research Anthology on Strategies for Using Social Media as a Service and Tool in Business, pp. 668–687. IGI Global (2021)

[21] Wright, C.I., Fischer, H., Whalen, P.J., McInerney, S.C., Shin, L.M., Rauch, S.L.: Differential prefrontal cortex and amygdala habituation to repeatedly presented emotional stimuli. Neuroreport **12**(2), 379–383 (2001)