Multimedia Tools and Applications (2024) 83:18779–18802
https://doi.org/10.1007/s11042-023-14704-7

A large-scale television advertising dataset for detailed impression analysis

Li Tao[1] · Shunsuke Nakamura[1] · Xueting Wang[2] · Tatsuya Kawahara[3] · Gen Tamura[3] · Toshihiko Yamasaki[1]

Received: 26 April 2021 / Revised: 29 April 2022 / Accepted: 4 February 2023 /
Published online: 1 August 2023

Abstract

Creating impressive video content such as movies and advertisements is a very important yet challenging task in business that requires both a sense of creativity and a lot of experience. Even professionals cannot necessarily invoke the impressions and emotions that they have aimed at. Many video advertisements are created and then disappear without giving a large impact on viewers. This paper presents a large-scale dataset of television (TV) advertisements that consists of 14,490 videos. The impressions of each video such as the recognition rate and interestingness rate are from the results of questionnaires answered by 620 participants. We also present a baseline for predicting the impression effects of TV advertisements by using visual and audio information, metadata such as broadcasting pattern, business category, the popularity of the casts, and text information including texts appearing on videos and narrations in audios. We predict four impressions of the viewers: 1) how much participants remember the video afterward, 2) how much they feel like buying the product/service, 3) how much they become interested in the product/service, and 4) how much they like the content of the advertisement itself. By combining images, audio, metadata, cast data, and text data, our baseline method is able to predict such impressions with a correlation of 0.69-0.82, much better than using a single-modal feature such as visual data or audio data only. This paper also gives some possible applications such as estimating the importance scores of each key frame, which gives us informative insights about how to make the advertisement content more impressive.

Keywords TV advertisement dataset · Multi-modal regression · Impression analysis · Impression prediction

1 Introduction

Predicting the effects of advertising is a critical issue for advertisers, but factors that contribute to the impression or emotion of viewers are still unknown. In TV advertisements,

✉ Toshihiko Yamasaki
yamasaki@cvm.t.u-tokyo.ac.jp

Extended author information available on the last page of the article.

the gross rating point (GRP) is a standard measure to estimate the effect. GRP is simply calculated by summing the viewer ratings of individual advertisements. GRP is believed to be correlated with the recognition rate (i.e., the percentage of people who remember the advertisement) because the frequency and the popularity of the TV program would be more correlated with the number of people who watch the advertisement and, in turn, recognize (remember) it. However, as we will show in Section 5, the correlation coefficient between the GRP and the actual recognition rate is quite small. Other factors such as how the advertisement will influence the viewers' desire to buy the products are much harder to predict because such impressions and emotions have little to do with GRP. To better understand the impacts on the audiences, we proposed new metrics to evaluate the impacts of an advertisement based on the real impressions from the audience.

The contributions of this paper are as follows. First, we construct a large-scale dataset of TV advertisements. The dataset contains 14,490 video clips with 23 kinds of subjective labels, which can be easily used in a supervised learning manner in the computer vision community. As far as we know, this is the largest TV advertisement video dataset with such rich labels. All the advertisement videos are created by professionals and actually broadcast on TV programs. The videos are evaluated by actual residents of the same country, not by crowd workers all over the world, which means these TV advertisements are promoted to the target audience without any cultural gap. This dataset can be shared with researchers in academia who are interested in proper contracts. Second, we present a multi-modal baseline which takes multi-modal features into consideration, analysis among each modality will also be discussed. In this paper, we tackle the task of predicting the following four impressional and emotional effects:

- Recognition rate: how much participants remember the advertisement.
- Willingness rate: how much participants feel like buying the product/service.
- Interestingness rate: how much participants become interested in the product/service.
- Favorableness rate: how much participants like the content of the advertisement itself.

These four dimensions out of 23 kinds of labels are much more significant according to the professionals from the advertisement industry. To the best of our knowledge, such high-level impressional effects and predictions for ad videos have never been tackled before.

Experimental results using 11,373 videos in our dataset show that our model can predict the impressional and emotional effects with correlation coefficients of 0.69-0.82, which is a vast improvement compared to the current de-facto standard of predicting using the GRP. When using individual models to predict each target, the results can be further improved to 0.69-0.85. Our experimental results show that the high-level impressional and emotional effects can be predicted with moderately high correlation coefficients, which we believe is beneficial for the advertisement industry and can be used as a baseline for future studies.

The remainder of this paper is organized as follows. Section 2 summarizes the related work in this research field. Section 3 describes our dataset. Our baseline of multi-modal network architecture is explained in Section 4 and the experimental results are shown in Section 5. Some of the possible applications are discussed in Sections 6 and 7 discusses limitations and future works. Concluding remarks are given in Section 8.

The techniques used in this paper might not novel, but the main scope of this paper is the world-largest TV advertisement dataset with detailed labels and metadata and a high baseline for such subjective effects.

2 Related work

2.1 Machine learning on advertisements

Hussain et al. [37] constructed a dataset describing what actions the viewer is prompted to take, the reasoning that the ad presents to persuade the viewer, and the symbolic references that the ad makes. The authors concluded that the prediction of topic and sentiment in video advertisements is still a very difficult problem. They only achieved 35.1% accuracy for predicting the video topic, and 32.8% accuracy for predicting the sentiment.

Impact analyses in advertisements have been studied in [76, 77]. In [76], immediate recall, day-after recall, and user experience are evaluated in addition to conventional effective measures such as valence and arousal. As compared to these works, the number of our raters is larger (620 vs 17 [76]), and much more detailed affective effects are labeled (23 vs 5 [76]) in our dataset. Adamov and Adali [2] employed sentiment analysis to find the most relevant advertisement to the main topic of the web page and the sentiment (positive or negative) of the author. This can be used for more context-aware advertisement, but [2] only analyzed the sentiment and not the effects of the content as we do in this paper.

One of the big issues of previous works is that the video dataset is collected from YouTube and the questions are answered by crowd workers. Therefore, the cultural background of the advertisements and viewers are not necessarily aligned. On the other hand, our study directly predicts the effects of the advertisements such as how much the consumers would be aware of the advertisements, how much the advertisements will succeed in persuading viewers to buy the products, and so on. This has been made possible by constructing a larger TV advertisement dataset and conducting extensive research on their impressional and emotional effects.

A lot of affective advertising generation methods have also been discussed. The authors in [95, 100] proposed video-in-video frameworks. In [95], object recognition and human detection are used for retrieving the optimal advertisement in terms of the advertisement's attractiveness and intrusiveness to the viewers. Zhang et al. [100] used eye-gaze tracking data to minimize the user disturbance because of advertisement insertion while enhancing the user engagement with the advertising content. Similar works aiming at text insertion onto videos can also be found in [53]. Yashima et al. [97] proposed a joint image and language DNN model to learn from the noisy online data and produce product descriptions that are closer to human-made descriptions with more subjective and impressive attribute expressions. They succeeded in generating more emotionally impressive descriptions according to the crowd workers. A similar study can be found in [56].

Advertisement video classification is also a fundamental problem of video categorization [36]. On the other hand, such category information can be used as one of the features.

Studies on infographics have also been conducted for designing more visually appealing poster advertisements. Saleh et al. [74] demonstrated that the combination of histograms of oriented gradients (HoG) [19] and color histograms was the most efficient method in retrieving similar style infographics, which can also be used for classification. Bylinskii et al. [10] proposed using neural networks trained on human clicks and importance annotations on hundreds of designs for predicting the relative importance of different elements in data visualizations and graphic designs. The prediction model of importance could be used for automatic design retargeting and thumbnailing [11]. However, it focused only on making more eye-catching static posters. In [102], a deep learning framework was proposed for exploring the effects of various design factors on perceived personalities of graphic designs and applications to element-level design suggestion and example-based personality transfer were demonstrated. Color enhancement for advertising images has also been discussed [14].

Recently, research on online banners are also reported [16, 28, 40, 65, 91, 92]. The first DNN-based CTR model was reported in [16]. In [28], the influence of the image ad's dimensions such as width and height on CTR were discussed. Xia *et al.* reported that the click through rate (CTR) for online image advertisements can be predicted with high accuracy [91, 92]. The correlation coefficient of the CTR prediction model was 0.828 using a multi-modal regression model. A similar idea of multi-modal fusion is also reported in [65]. The CTR prediction for videos was also achieved with the correlation coefficient of 0.70 [40] recently.

Producing suitable advertisements for the target audience is also close to the field of recommendation system, which also tries to show appealing information to users. Deep autoencoder-based methods [61, 62] can handle sparse data and produce good feature representations. Pan et al. [63] used a stacked denoising autoencoder to obtain compact representations. For sparse input data, graph convolution network (GCN) has also been applied and shown its effectiveness in [35, 98] in this research field.

2.2 Physiological analysis on advertisements

Multiple physiological studies on impressions and emotions invoked by advertisements have been conducted. Aaker et al. [1] analyzed the "informativeness" of 524 prime time TV advertisements that were rated by more than 250,000 viewers. The correlation between informativeness and other personal relevance adjectives suggested that an informative advertisement tends to be worth remembering, convincing, effective, and interesting. Vaughn [87] proposed a model based on a matrix of consumer thinking-feeling and high-low involvement behaviors. He suggested that some purchase decisions are well thought out whereas others are more impulsive, requiring less involvement on behalf of the consumer.

Yang et al. [96] demonstrated that users' zapping behavior has a close relationship with their facial expressions. Based on such investigations, an advertising evaluation metric, Zapping Index (ZI), was proposed to measure a user's zapping probability. ZI could also be used to measure a user's preference for different categories of advertisements, which would assist advertisers as well as publishers in understanding users' behavior.

The relationship between the higher-level impressions as we focus on in this paper and the eye-gaze [100] or multimodal features including the RR Intervals of heart beat [60] is also investigated.

Those technologies can be used for predicting the emotional effects of advertisements, but the experimental setup is troublesome. From that point of view, automatic prediction of effects without extra interaction with humans is preferred.

Our work is different from the above in that our DNN-based model can predict the effects even before actually broadcasting (or watching) the content or conducting a large-scale subjective study. Therefore, it would become possible to give insightful suggestions to the advertisement industry.

2.3 Emotion, Sentiment, and Memorability

Sentiment classification and effective analysis have been conducted for text [3, 6, 22, 39, 64], speech [31, 44, 57, 89, 101], audio [12, 80, 88], image [50, 52, 55, 66, 99], face [15, 17, 25, 72, 103], and video [7, 41, 58, 59, 94]. There are also large-scale datasets of images annotated with their emotion/sentiment such as the BAM! dataset [90], EMOTIC [49], SentiBank [9], and so forth. SentiBank is a large-scale Visual Sentiment Ontology (VSO) dataset consisting of more than 3,000 Adjective Noun Pairs (ANP). The authors also presented a visual concept detector library that can be used to detect the presence of 1,200 ANPs in an image.

Most of the works use Ekmans' atlas of emotions [23] or Plutchik's wheels of emotions [68]. Plutchik's wheels of emotions are similar to those of Russell's model in this regard and differ from Ekman's basic emotions. Plutchik's theory allows us to clearly perceive the proximity between arbitrary pairs of emotion categories. In either case, however, the target emotions are elemental ones such as joy, anger, and so on. Higher-level impressions such as those in our study have rarely been discussed so far.

Some researchers have discussed the memorability of images [5, 21, 26, 47, 67] and image/video manipulation techniques to enhance memorability [27, 78]. Kim et al. [48] showed that heart rate and galvanic skin response (GSR) are major predictors of memorability for photographers. In many cases, memorability is measured after showing the content once. On the other hand, we assume that the participants have watched the advertisements multiple times somewhere else before answering the questionnaires.

Compared to the memorability of images, the memorability of videos is still an emerging field of research [18, 30, 43, 75]. Han et al. [30] proposed a computational model that can correlate low-level audiovisual features with brain activity decoding using fMRI. Shekhar et al. [75] discussed efficient features for video memorability and applied it to video summarization.

On the other hand, the memorability of advertisements is complexly affected by their visual content, audio, narrations, GRP, and so on.

3 TV advertisement dataset

3.1 Data

As is well known, a large amount of data is required in order to properly train DNN models, but such datasets of TV commercials are not available as long as we know mainly because of copyright issues. We constructed a large-scale TV advertisement dataset that contains 14,490 video clips that were actually broadcast in Japan on TV from January 2006 to April 2016. All the advertisements are repeatedly broadcast on TV. Those broadcast only a few times (e.g., special advertisements dedicatedly designed for a big sports event) are not included. The quality of the archived videos is also carefully controlled: the advertisements are digitized by using the same series of analog-digital converters of the same brand for nine years in order to make the encoding quality deviation as small as possible. After the digital broadcasting service started, we used digital-analog converters to encode the videos. The resolution of the videos is 320 × 240. The number of videos per clip-length is summarized in Fig. 1. Examples of videos in our dataset are shown in Fig. 2. It can be noticed that most of the videos are 15 seconds, which is standard in the country.

There is also information on the advertiser, GRP, characters/animals/objects in the video, speech script, service category (e.g., food, car, etc.), and broadcasting pattern. There are eight types of broadcasting patterns. A typical pattern is broadcasting all day for the whole week, and another one is broadcasting only in the evening on weekdays and all day on weekends. Advertisements for compact or family cars usually take the first strategy and those for sports or luxury cars take the latter strategy based on the target consumers.

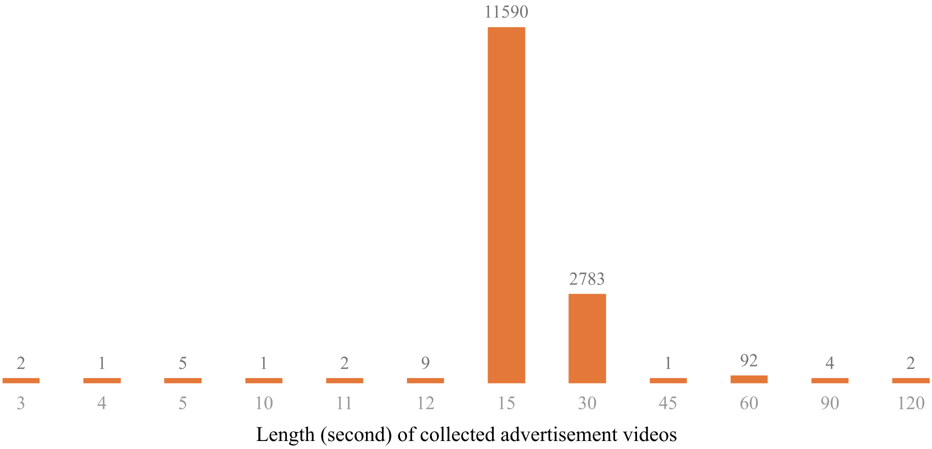

Fig. 1 The number of advertisement video clips in our database. The most of TV advertisement videos are 15 seconds, which is standard in the TV advertisement industry in Japan

There are 16 business categories: material, foods, medicine, cosmetics, fashion, publishing, industrial machines, office supplies, electronics, transportation, house supplies, houses, shops, finance, service/leisure, and others. All the mentioned information can be treated as metadata. And all these twenty-three kinds of metadata are provided as shown in Table 1.

The popularity of the casts in advertisements such as actors, actresses, singers, etc. is also an important factor that can influence the recognition rate and so on. Therefore, the recognition rate and popularity rate of famous talents were also investigated by asking a different set of 565 people twice a year.

Fig. 2 Sample images of TV advertisements in our dataset

Table 1 Available metadata

ID	Description
1	Date of the investigation
2	Age and gender category of participants (11 classes)
3	Subtitles
4	Speech scripts
5	Major business category (out of 16)
6	Intermediate-level business category
7	Detailed business category (out of 256)
8	Series or non-series
9	Advertiser
10	Name of the product/service and the series
11	Name of the characters
12	The number of participants
13	The number of participants who remember the ad.
14	The number of advertising at the point
15	Total seconds of advertising at the point
16	Estimated cover ratio of the targeted population
17	Estimated cover ratio of the targeted household
18	Starting date of the advertisement
19	Ending date of the advertisement
20	Broadcasting pattern (out of 8)
21	Individual-wise GRP
22	GRP
23	Duration of the advertisement

We generated the cast features as follows. We first extracted the recognition rate and the favorability for famous talent in the advertisement. The number of famous talents in the advertisement will be set as an input feature vector, together with the average popularity rate of casting famous talents. As a result, a two-dimension feature vector is obtained to represent cast features for each advertisement.

Advertisements contain visual and audio information. And text information exists in images/frames and audio. Though visual and audio feature extractors can extract features with semantic information, these features cannot include detailed descriptions because it is beyond the color combination or the rhythms of audio. And the text-related information should also play an important role in TV advertisements. Therefore, text information should be set as another important element in the advertisement dataset. Because text information exists in both frames and audio, we collected two kinds of the text information according to the source: texts in frames and narration data.

To summarize, our dataset contains 14,490 Japanese TV advertisements which were broadcast on TV repeatedly some time from 2006 to 2016. Each advertisement in the dataset not only includes the video and audio information, but also contains metadata information, the popularity of acting casts, and other related texts information. The impressions of a TV advertisement are based on different perceptions and should combine different data modalities, and we tried our best to make it informative.

3.2 Annotation

Each video was evaluated by a different set of around 620 participants, who can be divided into 11 categories according to their ages and genders. Eight out of eleven categories are males/females from 13 to 19, those from 20 to 34, those from 35-49, and those from 50-59. Note that males and females are in different categories. In addition, we also use three global categories to represent different interests of different groups, which are all males, all females, and all the ages/gender.

The questions are listed in Table 2. Participants answered each question on a different-point scale. The survey measures advertising recognition (question ID 1) using a three-point scale (3 = have seen it, 2 = probably have seen it, 1 = have not seen it), purchase intention (question ID 2) using a five-point scale (5 = want to buy very much, 4 = want to buy somewhat, 3 = neutral, 2 = do not want to buy very much, 1 = do not want to buy at all), advertising likability (question ID 3 and 4) using a five-point scale (5 = like very much, 4 = like somewhat, 3 = neutral, 2 = dislike somewhat, 1 = dislike very much). Other advertising perceptions used multiple-choice questions for survey.

We would like to note that this dataset has been collected by taking more than 10 years, and therefore the social conditions at the time, e.g., tie-up with famous music, the popularity of actors/actresses, etc., are also reflected in the dataset.

Table 2 Questions in the evaluation form

ID	Evaluation point
1	Do you remember/know the video?
2	Do you feel like buying the product/service?
3	Do you get interested in the product/service?
4	Do you like the content?
5	Is it fresh?
6	Is it impressive?
7	Is it unforgettable?
8	Is it commonplace?
9	Is it non heart touching?
10	Do you feel intimate?
11	Do you feel sympathy?
12	Is it emotional?
13	Do you feel alienation?
14	Is it boring?
15	Is it easy-to-understand?
16	Is it convincing?
17	Is it reliable?
18	Is it unconvincing?
19	Is it funny?
20	Is it unwearying?
21	Is it persistent?
22	Is it tiring?
23	Is it vulgar?

Our dataset is four times larger, more quality controlled, and contains much richer labels than the dataset presented in [38]. The other differences are summarized in Table 3. Data are available to researchers in academia under proper contract with the data provider.

4 Prediction of advertisement effects using attention-based multimodal framework: A baseline

Because we have 23 kinds of annotations on impressional/emotional effects in our dataset, we can totally have 23 individual prediction tasks. We consulted the professionals in the TV advertisement industry in Japan, and the first four tasks are considered the most valuable: recognition rate, willingness rate, interestingness rate, and favorableness rate.

Here we provide an **attention-based multimodal framework** to combine information of TV advertisement together for the prediction of the targets, in which way the model can focus on the important part for better inference.

4.1 Components in the Attention-based Multimodal framework

The network structure of our baseline is shown in Fig. 3. Six kinds of multimodal features are used in our model: visual, audio, metadata, cast data, and two kinds of text data, i.e, texts in frames and narration. Each of these parts vary from one to another because of different data formats and modalities. We apply different networks which have been validated effective for each data modality to cope with different parts. By taking advantage of attention mechanism, our framework can efficiently make use of different data. And the attention weights can indicate which part is more important, which can also inspire the TV advertising industry to generate more appealing advertisements.

For visual features, key frames are extracted every second and fed to a ResNet-50 [33]. The ResNet is pre-trained using ImageNet [71] and no fine-tuning is conducted because we think this is enough to extract necessary semantic features from images; ResNet is simply treated as an image feature extractor. We also think to take the chronological order of the frames into account using networks such as Recurrent Neural Networks (RNNs). However, for TV advertisements, the scenes are changing in seconds, which makes them more

Table 3 Comparison of the dataset

Factor	[38]	Ours
# of images	64,832	–
# of videos	3,477	14,490
# of business category	38	16
# of labels	30	23
# of participants	–	8.6M
# of participants/video	about 6	620
Source	YouTube	Ad. in a country
Participants from	all over the world	same country
Task	classification	regression
Labels	superficial	deeper

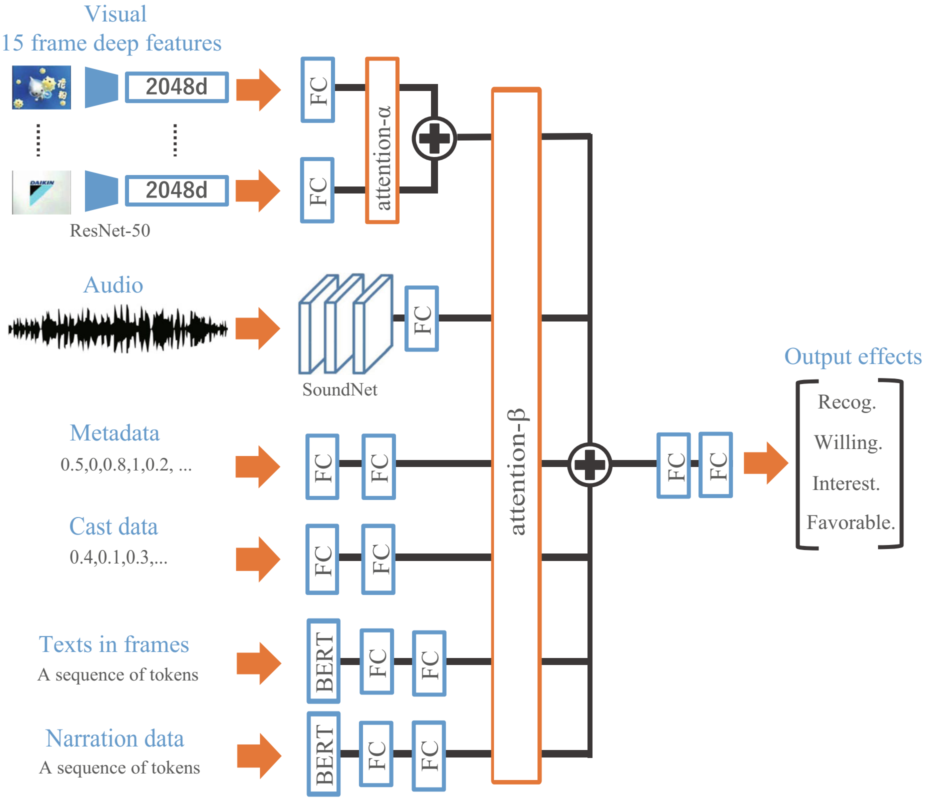

Fig. 3 Our DNN structure for advertisement effect prediction. Six kinds of input data are taken into consideration for final prediction

complex. A three-dimensional (3D) Convolutional Neural Network (3D-CNN) based network [42, 69, 83] was tried, but we confirmed that this approach did not work as well as our architecture in our preliminary experiments.

For the audio features, we employed SoundNet [4] because the network is designed for scene classification using sound, instead of human voice recognition. Although there are many other DNN-based and non-DNN-based approaches to sounds [12, 24, 80, 88], most of them were proposed to focus on human voice and did not work well for our purpose because advertisement videos usually contain not only human voices, but also music and sound effects. The SoundNet was firstly pre-trained using UrbanSound8K [73] and fine-tuned with our dataset.

The metadata and cast data have no spatial or temporal structure like video and audio, so they are directly fed to a multi-layer perception (MLP) with ReLU activation and dropout function.

Text information is in sentences in sequence. Traditional processing procedure split first split sentences into words and conduct word embedding to construct word vectors. Then Long Short-Term Memory (LSTM) or recurrent networks (RNNs) networks are will be used to encode word vectors to obtain the sentence vector. Recently, transformer [86] has drawn great attention in natural language processing. And Bidirectional Encoder Representations from Transformers (BERT) [20] has become one of the most famous network for word embedding. We make use of pretrained Japanese BERT model[1] to embed sentences and use the $[CLS]$ token as the sentence embedding for post-processing.

4.2 Two-step attention

The most important technical contribution in this paper is the two-step attention mechanisms, which are represented as α and β in Fig. 3: one for visual feature importance and the other for multimodal feature importance. Attention-α assigns weights to the key frames to conduct weighted-average pooling:

$$\mathbf{F}_{\text{frame}} = \sum_{i=1}^{15} \alpha_i f_i, \tag{1}$$

where $\mathbf{F}_{\text{frame}}$ is a visual feature of the input video and $\mathbf{f}_i$ is the feature extracted using ResNet for each frame. $\mathbf{F}_{\text{frame}}$ is calculated by (weighted) sum of key frame features rather than concatenating them, which is employed in literature [46, 54]. α is the attention vector indicating the importance of each frame, i.e., α_i denotes the attention weight for i_{th} frame. α is calculated by the following equation:

$$\alpha = \sigma_2 \mathbf{W}_2(\sigma 1 \mathbf{W}_1 \mathbf{f} + b_1) + b_2, \tag{2}$$

where σ_1 and σ_2 are sigmoid and softmax functions, respectively, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrices of the neural network. $\mathbf{f}$ is the concatenated feature from $[f_1, ..., f_{15}]$. b_1 and b_2 are the biases in the linear layer.

The other attention, attention-β, is used when combining the different multimodal features:

$$\begin{aligned}\mathbf{F}_{\text{multimodal}} = {} & \beta_1 \mathbf{F}_{\text{visual}} + \beta_2 \mathbf{F}_{\text{audio}} + \beta_3 \mathbf{F}_{\text{meta}} \\ & + \beta_4 \mathbf{F}_{\text{cast}} + \beta_5 \mathbf{F}_{\text{text}} + \beta_6 \mathbf{F}_{\text{narration}},\end{aligned} \tag{3}$$

where $\mathbf{F}_{\text{visual}}$, $\mathbf{F}_{\text{audio}}$, $\mathbf{F}_{\text{meta}}$, $\mathbf{F}_{\text{cast}}$, $\mathbf{F}_{\text{text}}$, and $\mathbf{F}_{\text{narration}}$ are deep features for visual, audio, metadata, cast data, text in frames, and narration data, respectively. And $\mathbf{F}_{\text{multimodal}}$ is the generated multimodal feature for the input video, which is fed to the full-connection (FC) layers for the advertisement effects prediction. β is the attention vector indicating the importance of each kind of feature. β is calculated in the same manner as α:

$$\beta = \sigma_2 \mathbf{W}_2 \left(\sigma 1 \mathbf{W}_1 \mathbf{F}^T + b_1\right) + b_2, \tag{4}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$, b_1 and b_2 are the weights and biases in the linear layer. Note that these values are from a different linear layer from those in (2), we do not use additional symbols to distinguish them. $\mathbf{F}$ is obtained by

$$\mathbf{F} = \mathbf{concat}(\mathbf{F}_{\text{visual}}, \mathbf{F}_{\text{audio}}, \mathbf{F}_{\text{meta}}, \mathbf{F}_{\text{cast}}, \mathbf{F}_{\text{text}}, \mathbf{F}_{\text{narration}}). \tag{5}$$

The four kinds of impressional and emotional effects listed in Section 1 (ID 1-4 in Table 2) which are represented as percentages are jointly trained and all the errors are backpropagated to these networks. Note that the ResNet-50 and BERT are used only as feature extractors without parameter changing during training. We also conducted experiments for each task individually, meaning only one model is trained for one target, which requires

[1] https://github.com/cl-tohoku/bert-japanese

N_{target} models for N_{target} targets. In this way, the model will focus on the targets individually. Therefore, for the targeted emotional effects, individual models can usually obtain better performance than using one model to predict all effects.

4.3 Detailed settings

For better understanding of the network architectures, we list the detailed information for each data part in Table 4, which is also corresponding to each part in Fig. 3.

For different parts, they are in different data modalities. Therefore, the input dimensions vary from one to another. For example, there are 15 frames in shape 224 × 224 × 3 for visual information, where 224 is the size of width and height and 3 represents the RGB channels; audios are converted into raw waveform and then sampled to a fixed dimensional (i.e., 661,500) vector using a global pooling strategy [85], as the original SoundNet [4] did; the dimension of metadata and cast data corresponds to the construction in our dataset; and for texts in frames and narrations, sentences are split into tokens and converted to token ids where the vocabulary size is 32,000, and the maximum sequence length is 512.

Many pre-trained models prove their effectiveness for feature representation. Therefore, in our work, we directly use some pre-trained models for pre-processing. ResNet-50 is used for frame feature extraction and BERT-base is used for sentence embedding, whose parameters are fixed without fine-tuning. The pre-trained SoundNet (5 Layer) model weights are also used for faster convergence. The dimensions of these models vary from one to another. Therefore, additional FC-BN-ReLU layers are necessary to convert them into the same shape.

Table 4 Details of each part in our solution. Texts in **bold** indicate that pre-trained models are used for feature extraction without parameter tuning. Texts in frames and narrations are under the same processing procedure, and we use Texts here to show the corresponding network architectures

Part	Network	Input dim	Output dim
Visual	**ResNet-50** [34]	$15 \times 224 \times 224 \times 3$	15×2048
	FC-BN-ReLU	15×2048	15×256
Audio	SoundNet [4]	$1 \times 1 \times 661500$	256
	FC-BN-ReLU	256	256
Metadata	FC-BN-ReLU	26	128
	FC-BN-ReLU	128	256
Cast data	FC-BN-ReLU	2	128
	FC-BN-ReLU	128	256
Texts	**BERT-base** [20]	32000×512	768
	FC-BN-ReLU	768	256
	FC-BN-ReLU	256	256
Attention	FC-LeakyReLU	$N \cdot 256$	256
	FC-Sigmoid	256	N
Prediction head	FC-BN-ReLU	256	256
	FC-Sigmoid	256	$N_{targets}$

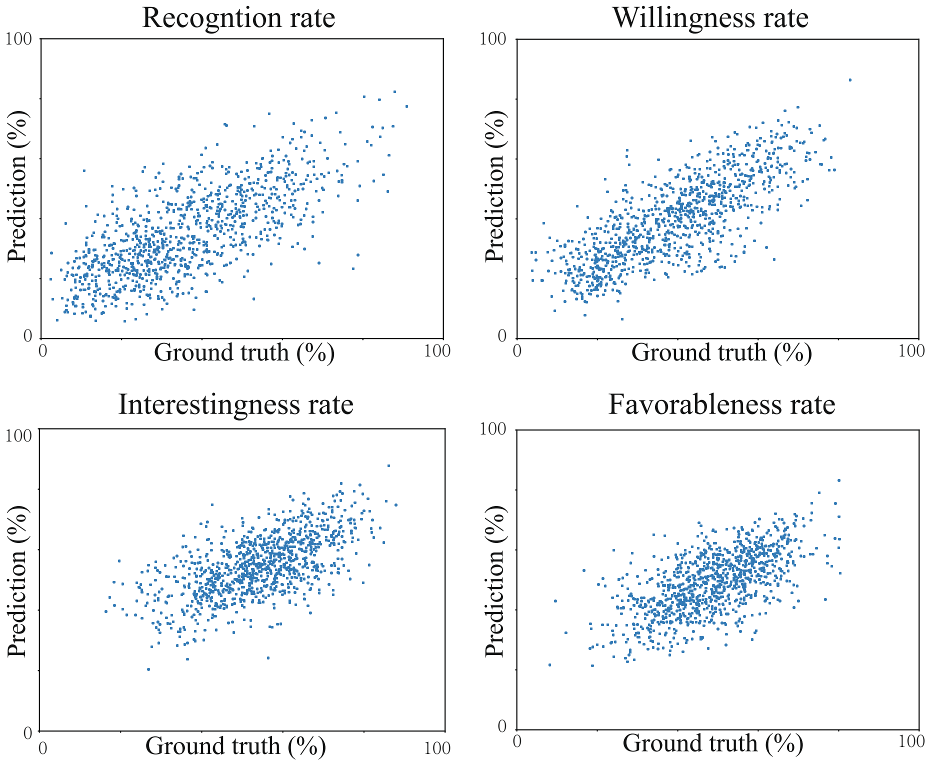

Fig. 4 Predicted and actual effects for our proposed model

In the two-step attention modules, attention weights will be generated using a sigmoid function for each branch. After aggregating all features, a projection head is used for the prediction of our targets.

During training, the batch size is set to 16. The initial learning rate is 0.1, and will multiply a decreasing factor every 5 epochs until it becomes 0.01 for the last epoch. The model is trained for 150 epochs. If not mentioned, the model will be trained to predict the recognition rate, willingness rate, interestingness rate, and favorableness rate at the same time. When training individual models, only one prediction is activated.

5 Experimental results

In the experiments, we used 11,373 video clips whose lengths were 15 seconds out of 14,490 videos in our dataset. We focused only on 15-second data because they were dominant. We eliminated two kinds of advertisements: informative advertisements that were not designed to sell products and those whose GRP was less than 10 (i.e., broadcast only a few times). The metadata employed for these 14,490 videos in our experiments were GRP, business category (i.e., 14 categories), series type (i.e., 4 types), and broadcasting pattern (i.e., 7 time slots), forming totally 26 dimensions for the input data. Namely, only the metadata that can be available before actually broadcasting were used in order to facilitate before-broadcast prediction. Images were sampled every second, and therefore 15 key frames were extracted from each video clip. The frames were resized to 224 × 224. The numbers of video clips used for training, validation, and testing were 9,373, 1,000, and 1,000, respectively.

The relationship between the predicted and actual recognition (R), willingness (W), interestingness (I), and favorableness (F) rates are shown in Fig. 4. As we mentioned in Section 3, we calculated the percentage of the participants who answered 1 (strongly yes) or 2 (weakly yes) to the questionnaire and regarded it as ground truth. Since the rates are predicted by regression, the agreement ratio among participants is not discussed in this paper. In addition, the correlation values and mean squared error (MSE) in terms of the percentage of the impressional/emotional effects are summarized in Table 5. We can observe that GRP is not a very good measure for predicting the impressional/emotional effects. Even the correlation value for the recognition rate, which seems somewhat correlated with GRP, is only 0.35. The other rates show only a non-important or very weak correlation with GRP. As a result, we have to say that there is little correlation between GRP and our target impressions.

The image features show the greatest predictive performance among the three individual features. We do not conduct experiments using cast data, texts in frames, and narrations individually because it is unreasonable to use cast data only when analyzing the impact of TV advertisements, and the text data here come together with visual/audio. Although each modality seems useless when it is used alone, the prediction accuracy is greatly improved when they are combined with the image features. Because we are coping with videos, in video understanding, there are many existing methods using 3D convolutional networks (3D CNN) to extract motion features [13, 32, 81–84, 93]. These methods were usually trained and tested in action recognition datasets [45, 51, 79]. We also applied C3D [83] to test its effectiveness in our TV advertisement dataset. The toy experiments only used visual, audio, and metadata. The results are in Table 6. We can find that better results are obtained with image-level features for visual embeddings. This may be caused by the complexity of TV

Table 5 Correlation coefficients and MSE of each approach on the test dataset. R, W, I, F represent rates for recognition, willingness, interestingness, and favorableness, respectively. For the individual model setting, one number represents the result from the corresponding model, and totally four models will be trained for R, W, I, F, respectively

Approach	Correlation ↑				MSE ↓			
	R	W	I	F	R	W	I	F
GRP	0.35	0.22	0.07	0.27	NA	NA	NA	NA
Visual (V)	0.47	0.58	0.50	0.48	273	76	127	114
Audio (A)	0.32	0.33	0.30	0.33	329	107	165	137
Metadata (M)	0.43	0.25	0.14	0.33	264	102	162	123
V + A	0.51	0.56	0.53	0.53	259	78	121	107
V + M	0.54	0.59	0.51	0.54	239	74	123	104
V + A + M	0.60	0.59	0.48	0.57	213	75	134	99
V + A + M + Cast (C)	0.69	0.72	0.62	0.72	169	67	99	52
All (V + A + M + C + Texts)	0.74	0.82	0.69	0.69	142	82	85	81
All (Individual model)	0.75	0.85	0.73	0.73	138	71	77	72

Table 6 Comparison between ResNet-50 (2D CNN) and C3D (3D CNN). R, W, I, F represent rates for recognition, willingness, interestingness, and favorableness, respectively

Approach	Correlation ↑				MSE ↓			
	R	W	I	F	R	W	I	F
V (2D CNN) + A + M	0.60	0.59	0.48	0.57	213	75	134	99
V (3D CNN) + A + M	0.57	0.51	0.37	0.52	221	81	146	102

advertisements because compared to the action recognition dataset, one TV advertisements contain many scenes and complex actions as well as human-object interactions, and the impression on the audience may also reply on the color, atmosphere, etc.

When we look back to Table 5, the best predictive performance is obtained when all features are used. Our proposed model can achieve remarkable improvement over a naive model which only uses a single modality. For one model which predicts four targets, the best correlation coefficients are 0.75, 0.85, 0.73, and 0.73, respectively. It is also interesting to see that predicting the interestingness rate is exceptionally difficult as compared to predicting recognition, willingness, and favorableness rates. This might be because that interestingness is judged by a different layer of emotions. By using our model, the prediction of willingness to buy and favorableness can become as accurate as that for the recognition whereas the GRP has little correlation with those factors. We also tried to train our models for each target task individually, reaching even higher performance (the last row in Table 5).

To better understand our proposed baseline prediction model, the attention weights used in our network are visualized in Figs. 5 and 6. Figure 5 shows the attention values (α) assigned to each frame of video for randomly sampled 10 videos. It can be observed the attention value is dynamically changing depending on the content. For the 4th, 5th, and the 10th samples, the attention weights paid for the first frames are very high. For the 1st,

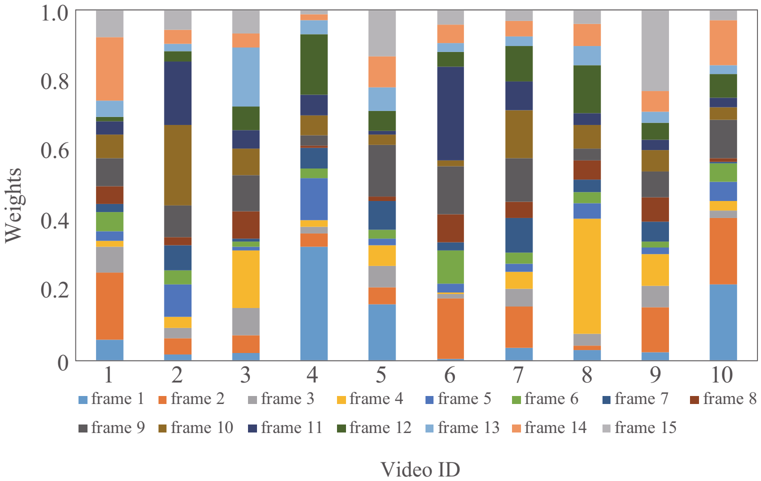

Fig. 5 Attention values for video frames. The importance for different frames vary from one to another. And our model can apply adaptive weights to each frame for better performance

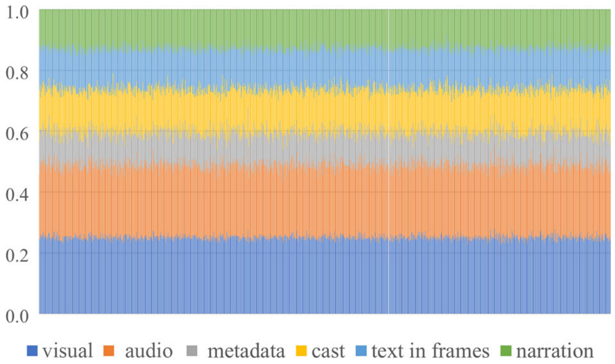

Fig. 6 Attention values for different data modalities. The most important features are visual and audio, which is similar to the common sense because TV advertisements are videos

5th, and the 9th samples, the attention weights paid for the last two frames are also apparently higher than average. These frames are usually where the product or the company logo appears. For other parts where the model pays more attention, such as frame 10 for 2rd sample and frame 4 for 8th sample, these situations are highly related to the specific content. Therefore, we can conclude that the first and last few parts in TV advertisement are comparably important. We also confirmed that our model pays attention only to a representative frame when a similar scene lasts for seconds.

The attention values for each data modality are shown in Fig. 6. As we can see from the figure, visual and audio play very important roles in advertisements because good vision or audio sound can attract the audience's attention at the very beginning of an advertisement. Texts in frames and narration data exist in both visual and audio modalities, which can not be ignored. Compared to metadata, cast data is more important, making it reasonable to pay high prices for superstars in advertising. We discussed with some professional creators of advertisements. They told us that they usually pay more attention to sound and music compared to visual content whereas movie creators had the opposite tendency. Therefore, this phenomenon coincides with the creators' common sense.

6 Applications

6.1 Online A/B testing

Advertisers usually create two or three versions of each advertisement and conduct A/B testing to decide which one to broadcast. A lot of people will be recruited to watch different versions and report their responses, and the final version will be chosen according to the responses. However, conducting such surveys is usually costly and time-consuming, with a risk of information leakage. Our system can predict recognition rate, willingness to buy, etc., giving advertisers a rough idea of how much impact each advertisement would have. An example is shown in Fig. 7, where two types of pet food advertisements are shown. Predicting which advertisements will be more remembered is a very difficult task, but our model successfully predicts the right one is more remembered than the left one.

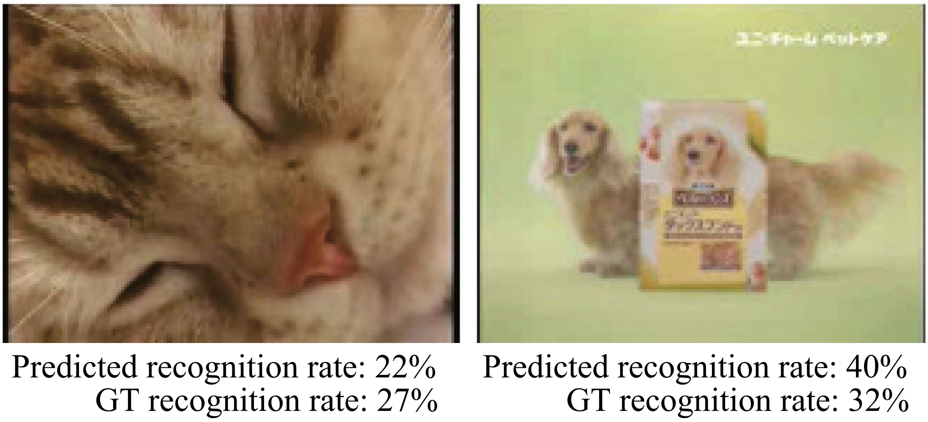
Predicted recognition rate: 22%
GT recognition rate: 27%
Predicted recognition rate: 40%
GT recognition rate: 32%

Fig. 7 Example of A/B testing. If two optional TV advertisements are provided, our model can predict the impression score of the given advertisements, which can help to choose better one for release

Creators would also benefit from our system by obtaining objective opinions when they have multiple ideas that they want to compare. Of course, this does not necessarily lead to only a single fully optimized advertisement because creators still have choices to make. And as a matter of course, the prediction accuracy cannot become 100%.

6.2 Scene factorization

The questionnaires asked questions about the entire video, not about each key frame. By introducing the technique in [8], we can estimate how much each scene or frame contributes to the impressional/emotional effect. The flow is illustrated in Fig. 8. By blackening one of the key frames and comparing the predicted impressional/emotional effects, the effect of the single key frame can be estimated. We denote the difference between the predicted value in the original video and the predicted value without a particular key frame as the importance score of the key frame. The calculated importance score of each key frame for the recognition rate is shown in Fig. 9.

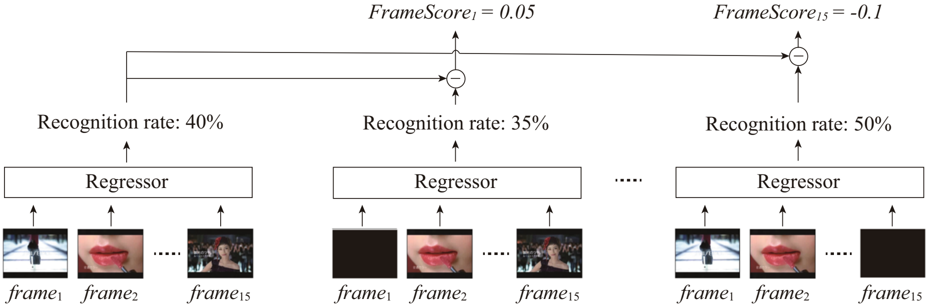

Fig. 8 Approach for scene factorization. This is only conducted based on frame-level regression

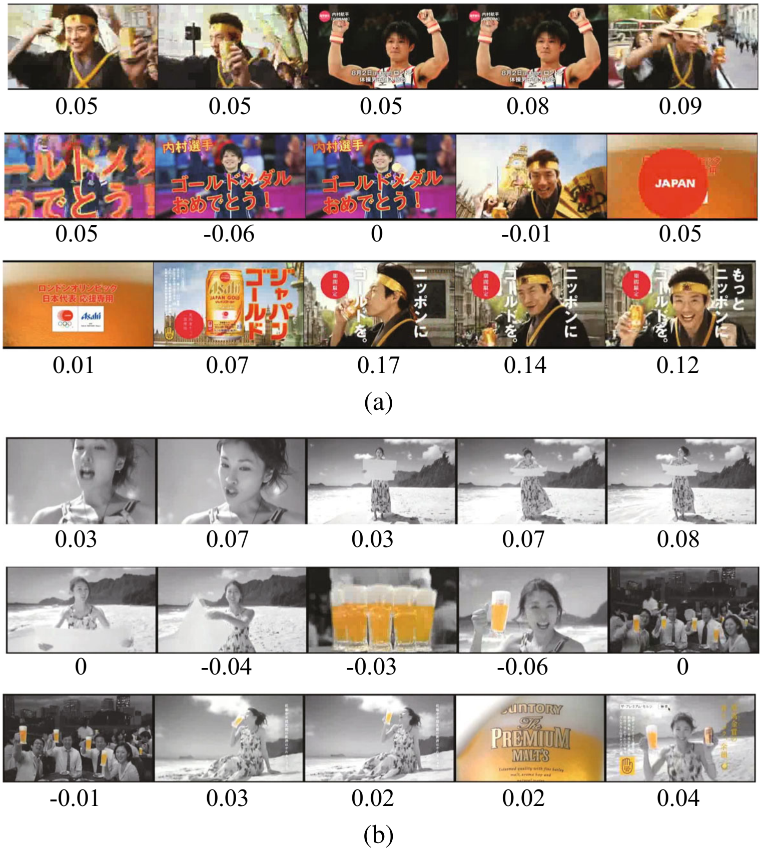

Fig. 9 Importance score for each key frame in terms of recognition rate. (a) advertisement of a canned coffee (predicted: 0.45, actual: 0.41) and (b) that of beer (predicted: 0.41, actual: 0.45)

Strictly speaking, our model calculates the estimated impressional/emotional effects for the case where a certain frame is replaced with a black scene, not for the case where the frame did not exist. Therefore, the absolute score is not reliable but the relative scores and their sign (negative or positive) can give us some insights. For instance, the existence of actors or actresses is said to be important for the advertisement to be remembered [29], and a similar tendency can be observed in Fig. 9. The scenes where the casts appear to yield higher scores for recognition. It is interesting to see that our DNN model can produce similar insights to conventional rules of thumb that are shared among professional creators. In addition, the company logo or the zoom shot of the product in the last scene gives a relatively large impact. Our system also detected scenes that might negatively contribute to the recognition rate as shown in Fig. 9. Note that the scores are relative importance values and the sum of the scores does not match the predicted final score.

7 Limitations and future work

The ads are all broadcast on TV and the participants are supposed to have watched them several times before the experiments. This is why we can ask them whether they have watched (or they remember) the ads. In other words, our dataset is not designed for before/after comparison. Investigation of such short-term effects is our future work.

We expect our model to be able to properly evaluate new-style advertisements because it is a learning-based approach. An algorithm that can evaluate novelty or creativity [70] might be needed.

We would also like to work on the optimization of advertisements and the creation of new advertisements using our proposed model. There are many scenes in one TV advertisement and one scene may not be sampled if it lasts less than 1 second in the current solution, and scene changes may also need to be considered.

8 Conclusions

In this paper, we presented a new dataset of TV advertisements with 23 kinds of annotations on impressional/emotional effects, which we believe is the largest and the richest data on impressional/emotional effects of advertisements. Our experiments using 11,373 15-second video clips showed that by combining visual data, audio data, metadata, cast data, and texts, we can predict impressional/emotional effects such as the willingness to buy the product in the advertisement with a correlation coefficient of 0.84. We showed some applications of our regression model.

We believe that our paper can contribute to the multimedia as well as the computer vision community by providing a large impression-related dataset of video and some baseline results. Furthermore, our technology can give creators insightful information to assist them in their creative work.

Acknowledgements This work was partially financially supported by the Grants-in-Aid for Scientific Research Numbers JP19K20289 and JP18H03339 from JSPS.

Funding Open access funding provided by The University of Tokyo.

Declarations

Conflict of Interests The authors have no conflict of interest to declare that are relevant to the content of this article.

References

1. Aaker DA, Norris D (1982) Characteristics of tv commercials perceived as informative. J Advert Res
2. Adamov AZ, Adali E (2016) Opinion mining and sentiment analysis for contextual online-advertisement. In: 2016 IEEE 10Th international conference on application of information and communication technologies (AICT), IEEE, pp 1–3
3. Al-Moslmi T, Omar N, Abdullah S, Albared M (2017) Approaches to cross-domain sentiment analysis: a systematic literature review. IEEE Access 5:16173–16192

4. Aytar Y, Vondrick C, Torralba A (2016) Soundnet: Learning sound representations from unlabeled video. Adv Neural Inf Process Syst 29
5. Bainbridge WA, Dilks DD, Oliva A (2017) Memorability: a stimulus-driven perceptual neural signature distinctive from memory. Neuroimage 149(Supplement C):141–152
6. Bakshi RK, Kaur N, Kaur R, Kaur G (2016) Opinion mining and sentiment analysis. In: 2016 3rd international conference on computing for sustainable global development (INDIACom), pp 452–455
7. Baveye Y, Chamaret C, Dellandréa E, Chen L (2017) Affective video content analysis: a multidisciplinary insight. IEEE Trans Affect Comput 9(4):396–409
8. Bazzani L, Bergamo A, Anguelov D, Torresani L (2016) Self-taught object localization with deep networks. In: 2016 IEEE Winter conference on applications of computer vision (WACV), IEEE, pp 1–9
9. Borth D, Ji R, Chen T, Breuel T, Chang SF (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM international conference on Multimedia, pp 223–232
10. Bylinskii Z, Kim NW, O'Donovan P, Alsheikh S, Madan S, Pfister H, Durand F, Russell B, Hertzmann A (2017) Learning visual importance for graphic designs and data visualizations. In: Proceedings of the 30th Annual ACM symposium on user interface software and technology, pp 57–69
11. Bylinskii Z, Kim NW, O'Donovan P, Alsheikh S, Madan S, Pfister H, Durand F, Russell B, Hertzmann A (2017) Learning visual importance for graphic designs and data visualizations. In: Proceedings of the 30th Annual ACM symposium on user interface software and technology, pp 57–69
12. Cakir E, Heittola T, Huttunen H, Virtanen T (2015) Polyphonic sound event detection using multi label deep neural networks. In: 2015 International joint conference on neural networks (IJCNN), IEEE, pp 1–7
13. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR), pp 6299–6308
14. Chae Y, Nakazawa M, Stenger B (2018) Enhancing product images for click-through rate improvement. In: 2018 25Th IEEE international conference on image processing (ICIP), IEEE, pp 1428–1432
15. Chen H, Li J, Zhang F, Li Y, Wang H (2015) 3d model-based continuous emotion recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1836–1845
16. Chen J, Sun B, Li H, Lu H, Hua XS (2016) Deep ctr prediction in display advertising. In: Proceedings of the 24th ACM international conference on multimedia, p 811–820
17. Chiranjeevi P, Gopalakrishnan V, Moogi P (2015) Neutral face classification using personalized appearance models for fast and robust emotion detection. IEEE Trans Image Process 24(9):2701–2711
18. Cohendet R, Demarty CH, Duong NQK, Engilberge M (2019) Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV)
19. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE Computer society conference on computer vision and pattern recognition, vol. 1. IEEE, pp 886–893
20. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies (naacl-hlt), association for computational linguistics, pp 4171–4186
21. Dubey R, Peterson J, Khosla A, Yang MH, Ghanem B (2015) What makes an object memorable? In: Proceedings of the ieee international conference on computer vision, pp 1089–1097
22. Ebrahimi M, Yazdavar AH, Sheth A (2017) Challenges of sentiment analysis for dynamic events. IEEE Intell Syst 32(5):70–75
23. Ekman P, Friesen WV, O'sullivan M, Chan A, Diacoyanni-Tarlatzis I, Heider K et al (1987) Universals and cultural differences in the judgments of facial expressions of emotion. In: Journal of personality and social psychology, vol 53. American psychological association
24. Eyben F, Weninger F, Gross F, Schuller B (2013) Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia, pp 835–838
25. Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM (2016) Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5562–5570
26. Fajtl J, Argyriou V, Monekosso D, Remagnino P (2018) Amnet: Memorability estimation with attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6363–6372
27. Fei M, Jiang W, Mao W (2017) Creating memorable video summaries that satisfy the user fs intention for taking the videos. Neurocomputing

28. Fire M, Schler J (2017) Exploring online ad images using a deep convolutional neural network approach. In: 2017 IEEE International conference on internet of things (ithings) and IEEE green computing and communications (greencom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (smartdata)
29. Gregov J (2008) The impact of actor presence in product placements on brand and ad attitudes and purchase intention, and the role of recognition and recall. In: Annual meeting of the NCA 94th annual convention
30. Han J, Chen C, Shao L, Hu X, Han J, Liu T (2014) Learning computational models of video memorability from fmri brain imaging. IEEE transactions on cybernetics 45(8):1692–1703
31. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E et al (2014) Deep speech:, Scaling up end-to-end speech recognition. arXiv:1412.5567
32. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d CNNS retrace the history of 2d CNNS and imagenet. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, pp 18–22
33. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
35. He X, Deng K, Wang X, Li Y, Zhang Y, Wang M (2020) Lightgcn: Simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pp 639–648
36. Hou S, Chen L, Tao D, Zhou S, Liu W, Zheng Y (2017) Multi-layer multi-view topic model for classifying advertising video. Pattern Recogn 68:66–81
37. Hussain Z, Zhang M, Zhang X, Ye K, Thomas C, Agha Z, Ong N, Kovashka A (2017) Automatic understanding of image and video advertisements. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1705–1715
38. Hussain Z, Zhang M, Zhang X, Ye K, Thomas C, Agha Z, Ong N, Kovashka A (2017) Automatic understanding of image and video advertisements. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1705–1715
39. Hutto C, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media, vol 8
40. Ikeda J, Seshime H, Wang X, Yamasaki T (2021) Predicting online video advertising effects with multimodal deep learning. In: International conference on pattern recognition, pp 2995–3002
41. Irie G, Satou T, Kojima A, Yamasaki T, Aizawa K (2010) Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. IEEE Trans Multimedia 12(6):523–535
42. Ji S, Xu W, Yang M (2012) Yu, k.: 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
43. Kar A, Mavin P, Ghaturle Y, Vani M (2017) What makes a video memorable? In: 2017 IEEE International conference on data science and advanced analytics (DSAA), pp 373–381, https://doi.org/10.1109/DSAA.2017.37
44. Kaushik L, Sangwan A, Hansen JH (2017) Automatic sentiment detection in naturalistic audio. IEEE/ACM Trans Audio Speech Lang Process 25(8):1668–1679
45. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S et al (2017) The kinetics human action video dataset. arXiv:1705.06950 1–22
46. Kaya H, Gürpınar F, Salah AA (2017) Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image Vision Comput 65:66–75
47. Khosla A, Raju AS, Torralba A, Oliva A (2015) Understanding and predicting image memorability at a large scale. In: Proceedings of the IEEE international conference on computer vision, pp 2390–2398
48. Kim S, Patra KAE, Kim A, Lee KP, Segev A, Lee U (2017) Sensors know which photos are memorable. In: Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems, pp 2706–2713
49. Kosti R, Alvarez JM, Recasens A, Lapedriza A (2017) Emotic: Emotions in context dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 61–69
50. Kosti R, Alvarez JM, Recasens A, Lapedriza A (2017) Emotion recognition in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1667–1675
51. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: a large video database for human motion recognition. In: International conference on computer vision (ICCV), IEEE, pp 2556–2563
52. Li T, Ni B, Xu M, Wang M, Gao Q, Yan S (2015) Data-driven affective filtering for images and videos. IEEE Trans Cybern 45(10):2336–2349

53. Liang Y, Liu W, Liu K, Ma H (2018) Automatic generation of textual advertisement for video advertising. In: 2018 IEEE Fourth international conference on multimedia big data (bigMM), IEEE, pp 1–5
54. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. Adv Neural Inf Process Syst 29:289–297
55. Lu X, Lin Z, Jin H, Yang J, Wang JZ (2014) Rapid: Rating pictorial aesthetics using deep learning. In: Proceedings of the 22nd ACM international conference on Multimedia, pp 457–466
56. Mathews A, Xie L, He X (2016) Senticap: Generating image descriptions with sentiments. In: Proceedings of the AAAI conference on artificial intelligence, vol 30
57. Mencattini A, Martinelli E, Ringeval F, Schuller B, Di Natale C (2016) Continuous estimation of emotions in speech by dynamic cooperative speaker models. IEEE Trans Affect Comput 8(3):314–327
58. Noroozi F, Marjanovic M, Njegus A, Escalera S, Anbarjafari G (2017) Audio-visual emotion recognition in video clips. IEEE Trans Affect Comput 10(1):60–75
59. Nwana AO, Avestimehr S, Chen T (2013) A latent social approach to youtube popularity prediction. In: 2013 IEEE Global communications conference (GLOBECOM), IEEE, pp 3138–3144
60. Okada G, Yonezawa T, Kurita K, Tsumura N (2018) Monitoring emotion by remote measurement of physiological signals using an rgb camera. ITE Trans Media Technol Appl 6(1):131–137
61. Pan Y, He F, Yu H (2019) A novel enhanced collaborative autoencoder with knowledge distillation for top-n recommender systems. Neurocomputing 332:137–148
62. Pan Y, He F, Yu H (2020) A correlative denoising autoencoder to model social influence for top-n recommender system. Front Comput Sci 14(3):1–13
63. Pan Y, He F, Yu H (2020) Learning social representations with deep autoencoder for recommender system. World Wide Web 23(4):2259–2279
64. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing, pp 79–86
65. Park KW, Ha JW, Lee J, Kwon S, Kim KM, Zhang BT (2021) M2fn: Multi-step modality fusion for advertisement image assessment. Appl Soft Comput 107116:103
66. Peng KC, Chen T, Sadovnik A, Gallagher AC (2015) A mixed bag of emotions: Model, predict, and transfer emotion distributions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 860–868
67. Perera S, Tal A, Zelnik-Manor L (2019) Is image memorability prediction solved? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops
68. Plutchik R (2001) The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci 89(4):344–350
69. Qiu Z, Yao T, Mei T (2017) Learning spatio-temporal representation with pseudo-3d residual networks. In: Proceedings of the IEEE international conference on computer vision, pp 5533–5541
70. Redi M, O'Hare N, Schifanella R, Trevisiol M, Jaimes A (2014) 6 seconds of sound and vision: Creativity in micro-videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4272–4279
71. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
72. Ryu B, Rivera AR, Kim J, Chae O (2017) Local directional ternary pattern for facial expression recognition. IEEE Trans Image Process 26(12):6006–6018
73. Salamon J, Jacoby C, Bello JP (2014) A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia, pp 1041–1044
74. Saleh B, Dontcheva M, Hertzmann A, Liu Z (2015) Learning style similarity for searching infographics. In: Proceedings of the 41st graphics interface conference, canadian information processing society, pp 59–64
75. Shekhar S, Singal D, Singh H, Kedia M, Shetty A (2017) Show and recall: Learning what makes videos memorable. In: Proceedings of the IEEE international conference on computer vision workshops, pp 2730–2739
76. Shukla A, Gullapuram SS, Katti H, Yadati K, Kankanhalli M, Subramanian R (2017) Affect recognition in ads with application to computational advertising. In: Proceedings of the 25th ACM international conference on Multimedia, pp 1148–1156
77. Shukla A, Gullapuram SS, Katti H, Yadati K, Kankanhalli M, Subramanian R (2017) Evaluating content-centric vs. user-centric ad affect recognition. In: Proceedings of the 19th ACM international conference on multimodal interaction, pp 402–410

78. Siarohin A, Zen G, Majtanovic C, Alameda-Pineda X, Ricci E, Sebe N (2017) How to make an image more memorable? a deep style transfer approach proceedings of the 2017 ACM on international conference on Multimedia Retrieval, pp 322-329. Association for Computing Machinery, New York
79. Soomro K, Zamir AR, Shah M (2012). arXiv: preprint1212.0402 pp 1–7
80. Takahashi N, Gygli M, Van Gool L (2017) Aenet: Learning deep audio features for video analysis. IEEE Trans Multimedia 20(3):513–524
81. Tao L, Wang X, Yamasaki T (2020) Motion representation using residual frames with 3d CNN. In: 2020 IEEE International conference on image processing (ICIP), pp 1786–1790, https://doi.org/10.1109/ICIP40778.2020.9191133
82. Tao L, Wang X, Yamasaki T (2021) Rethinking motion representation: Residual frames with 3d convnets. IEEE Trans Image Process 30:9231–9244. https://doi.org/10.1109/TIP.2021.3124156
83. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
84. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6450–6459
85. Van den Oord A, Dieleman S, Schrauwen B (2013) Deep content-based music recommendation. Adv Neural Inf Process Sys 26
86. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
87. Vaughn R (1980) How advertising works: A planning model Advertising research
88. Wang JC, Lee YS, Chin YH, Chen YR, Hsieh WC (2015) Hierarchical dirichlet process mixture model for music emotion recognition. IEEE Trans Affect Comput 6(3):261–271
89. Wang K, An N, Li BN, Zhang Y, Li L (2015) Speech emotion recognition using fourier parameters. IEEE Trans Affect Comput 6(1):69–75
90. Wilber MJ, Fang C, Jin H, Hertzmann A, Collomosse J, Belongie S (2017) Bam! the behance artistic media dataset for recognition beyond photography. In: Proceedings of the IEEE international conference on computer vision, pp 1202–1211
91. Xia B, Seshime H, Wang X, Yamasaki T (2020) Click-through rate prediction of online banners featuring multimodal analysis. Int J Semant Comput 14(1):71–91
92. Xia B, Wang X, Yamasaki T, Aizawa K, Seshime H (2019) Deep neural network-based click-through rate prediction using multimodal features of online banners. In: IEEE International conference on multimedia big data, pp 162–170
93. Xie S, Sun C, Huang J, Tu Z, Murphy K (2018) Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the european conference on computer vision (ECCV), pp 305–321
94. Xu B, Fu Y, Jiang YG, Li B, Sigal L (2016) Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. IEEE Trans Affect Comput 9(2):255–270
95. Yadati K, Katti H, Kankanhalli M (2013) Cavva: Computational affective video-in-video advertising. IEEE Trans Multimed 16(1):15–23
96. Yang S, Kafai M, An L, Bhanu B (2014) Zapping index: using smile to measure advertisement zapping likelihood. IEEE Trans Affect Comput 5(4):432–444
97. Yashima T, Okazaki N, Inui K, Yamaguchi K, Okatani T (2016) Learning to describe e-commerce images from noisy online data. In: Asian conference on computer vision, Springer, pp 85–100
98. Ying R, He R, Chen K, Eksombatchai P, Hamilton WL, Leskovec J (2018) Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 974–983
99. You Q, Jin H, Luo J (2017) Visual sentiment analysis by attending on local image regions. In: Proceedings of the AAAI conference on artificial intelligence, vol 31
100. Zhang H, Cao X, Ho JK, Chow TW (2016) Object-level video advertising: an optimization framework. IEEE Trans Ind Inform 13(2):520–531
101. Zhang S, Zhang S, Huang T, Gao W (2017) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans Multimedia 20(6):1576–1590
102. Zhao N, Cao Y, Lau RW (2018) What characterizes personalities of graphic designs? ACM Trans Graph (TOG) 37(4):1–15
103. Zhen Q, Huang D, Drira H, Amor BB, Wang Y, Daoudi M (2017) Magnifying subtle facial motions for effective 4d expression recognition. IEEE Trans Affect Comput 10(4):524–536

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Li Tao[1] · Shunsuke Nakamura[1] · Xueting Wang[2] · Tatsuya Kawahara[3] · Gen Tamura[3] · Toshihiko Yamasaki[1]

Li Tao
chntaoli@gmail.com

Shunsuke Nakamura
kreski.linux.sea@gmail.com

Xueting Wang
wangxueting12@gmail.com

Tatsuya Kawahara
tatsuya.kawahara@videor.co.jp

Gen Tamura
gen.tamura@videor.co.jp

[1] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

[2] CyberAgent AI Lab, Tokyo, Japan

[3] Video Research Ltd., Tokyo, Japan