

Decoding Viewer Emotions in Video Ads: Predictive Insights through Deep Learning

Alexey Antonov

University of Warwick

Shravan Sampath Kumar

System1 Research

William Headley

System1 Research

Orlando Wood

System1 Research

Giovanni Montana

`g.montana@warwick.ac.uk`

University of Warwick

Article

Keywords:

Posted Date: November 9th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3553507/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on November 2nd, 2024.
See the published version at <https://doi.org/10.1038/s41598-024-76968-9>.

Decoding Viewer Emotions in Video Ads: Predictive Insights through Deep Learning

Alexey Antonov², Shravan Sampath Kumar⁴, William Headley⁴, Orlando Wood⁴, and Giovanni Montana^{1,2,3,*}

¹Department of Statistics, University of Warwick

²WMG, University of Warwick

³Alan Turing Institute

⁴System1 Research

*g.montana@warwick.ac.uk

ABSTRACT

Understanding and predicting viewers' emotional responses to videos has emerged as a pivotal challenge due to its multifaceted applications in video indexing, summarization, personalized content recommendation, and effective advertisement design. A major roadblock in this domain has been the lack of expansive datasets with videos paired with viewer-reported emotional annotations. We address this challenge by employing a deep learning methodology trained on a dataset derived from the application of System1's proprietary methodologies on over 30,000 real video advertisements, each annotated by an average of 75 viewers. This equates to over 2.3 million emotional annotations across eight distinct categories: anger, contempt, disgust, fear, happiness, sadness, surprise, and neutral, coupled with the temporal onset of these emotions. Leveraging 5-second video clips, our approach aims to capture pronounced emotional responses. Our convolutional neural network, which integrates both video and audio data, predicts salient 5-second emotional clips with an average balanced accuracy of 43.6%, and shows particularly high performance for detecting happiness (55.8%) and sadness (60.2%). When applied to full advertisements, our model achieves a strong average AUC of 75% in determining emotional undertones. To facilitate further research, our trained networks are freely available upon request for research purposes. This work not only overcomes previous data limitations but also provides an accurate deep learning solution for video emotion understanding.

Introduction

The exploration of emotions, particularly the endeavor to understand and interpret them, intersects the domains of psychology, technology, and artificial intelligence. Historically, systems aimed at discerning emotions have depended on overt and tangible cues. Facial expressions provide insights through delicate twitches of muscles that hint at concealed emotions. Voice analysis leverages slight variations in pitch and tone to uncover underlying sentiments¹. As technology advanced, researchers began employing more subtle cues through eye movement analyses, assessing pupil dilation and gaze direction to infer reactions like engagement or surprise. Delving even deeper into the neural realm, some pioneering studies have explored gleaning emotion-centric information directly from brain signals. While these approaches have proven fruitful, they are fundamentally limited to visible reactions and human interpretation.

Emotions constitute a profound human experience, shaped by intricate cultural norms, psychological states and physiological responses. Their diverse expressions depend heavily on subtle social context and individual differences. Accurately deciphering emotions' nuanced complexities thus presents deep challenges. However, effectively perceiving and interpreting human emotions could enable transformative applications in advertising and marketing. Granular insight into audience emotional resonance could allow for highly tailored, personalized ad campaigns and content. Identifying precise emotional reactions could inform data-driven targeting strategies, mood-based recommendation systems, uncovering unstated market research feelings, and crafting impactful creative narratives. The possibilities are extensive for generating more poignant, meaningful ads by deeply understanding emotions. Yet major obstacles remain in accurately recording diverse human emotional responses to multimedia content—a fundamental challenge requiring interdisciplinary breakthroughs.

Most prior emotion recognition has required direct human observation and annotation, introducing subjectivity and bias. Our study pioneers a novel perspective - predicting emotions purely from multimedia content, without any human input. We develop a deep learning system that annotates videos by hypothesizing the viewer emotions they may elicit, based solely on analyzing the audiovisual signals, devoid of direct human cues². This data-driven approach opens promising opportunities to understand emotions evoked by media in a more objective, scalable way. Potential applications are far-reaching, from enhancing recommendation engines to demystifying psychological aspects of filmmaking. Particularly promising is advertising,

where gauging emotional resonance could transform campaign strategies and creative testing.

The examination of emotions in academia is primarily anchored in two key models. The categorical model, founded on Ekman's pioneering work, asserts that emotions are distinct entities defined by specific expressive and physiological markers³. Through extensive cross-cultural studies, Ekman emphasized the universal nature of certain emotional expressions across diverse populations. This provides a robust foundation for analysis and wide-reaching applications like ours. Subsequent research has further refined categorical models, introducing more nuanced labels. Alternatively, the continuous model, exemplified by the circumplex of affect, envisions emotions on a fluid plane defined by valence and arousal⁴. This flexible perspective celebrates the intricacies and relationships between emotions. While the continuous view offers valuable insights into the subtle nature of emotions, we align here with Ekman's categorical approach for its structure and systematic framework. Clear distinctions between discrete emotions facilitate more straightforward mapping and prediction, streamlining the training process for machine learning.

Our study builds on extensive prior research, leveraging an exceptional proprietary dataset from System1 Group PLC, a pioneering leader in advertising effectiveness prediction. This dataset comprises over 30,000 video advertisements, each analyzed by approximately 75 diverse viewers who provided real-time emotional annotations. This equates to over 2.3 million rich labels across eight categories, enabling us to track nuanced temporal emotion shifts often hidden in traditional feedback. We develop a tailored convolutional neural network that ingests video and audio to detect prominent 5-second emotional clips and predict emotions in full ads. Our model achieves a mean balanced accuracy of 43.6%, with strong performance for detecting happiness (55.8%) and sadness (60.2%). On full videos, we obtain an impressive average AUC of 75%, demonstrating robustness despite the inherent subjectivity of emotion.

Despite this initial progress, substantial challenges remain in collecting diverse and unbiased emotional data, and developing algorithms that can learn conceptual connections between raw multimedia and complex psychological states. By publicly releasing our model and dataset, we aim to spur community efforts to advance emotion understanding in video analytics and empathetic AI systems. Moving forward, interdisciplinary collaboration and innovations in multimodal affective computing will be key. Overall, our work reveals promising steps in a multifaceted domain, catalyzing future research into the intricacies of human emotion and their machine interpretation.

Methods

Dataset

In our study, we made use of the "Test Your Ad" dataset, which was carefully curated by System1 Group PLC. This dataset comprises 30,751 video advertisements representing various commercial products. These ads were aired between 2017 and 2020 and targeted audiences in both the UK and US. What makes this dataset valuable is that the ads adhere to industry-standard durations, often set at intervals of 5, 15, 30, or 60 seconds. A breakdown of these durations can be found in Table 1.

The annotation process for these video ads was semi-automated and involved a panel of human reviewers. To aid this review process, an intuitive interface called FaceTrace® was used. Viewers could pause the videos at any point and select an image that best represented the emotion they were feeling. The interface continuously displayed a range of facial expressions, helping viewers align their emotions with the corresponding image during video playback. Figure 1 displays the eight facial expressions used in this process, encompassing Ekman's seven foundational emotions³: anger, contempt, disgust, fear, happiness, sadness, and surprise. Additionally, viewers could choose a neutral expression, although it wasn't consistently selected. This allowed viewers to indicate when a previously felt emotion had subsided.

The continuous visual cues on the interface enabled viewers to effortlessly connect their immediate emotional reactions with the matching facial expression. This eliminated the need for viewers to recall their emotions whilst watching the video. The result of this real-time annotation approach was a sequence of clicks for each video and viewer, each linked to a specific time point and associated with the relevant emotion. We refer to this series of temporal emotional annotations as the *emotional profile*. Remarkably, each video received annotations from an average of 75 viewers, amassing a total of 2.3 million emotional profiles. Upon closer examination, we found that there were 1.6 million unique completed reviews. In 99.8% of cases, an individual reviewer assessed only one or two distinct video advertisements.

Quantification of ground-truth emotional responses

In our endeavor to trace the nuanced shifts in viewers' emotional experiences while viewing videos, we introduce the term *emotional jump*. This term captures a quantifiable shift in emotional intensity that viewers undergo during a brief segment of a video. Depending on the content and viewer interpretation, the intensity of this jump might either escalate or diminish. Our methodology comprises two phases. Initially, we identify patterns of these emotional jumps across all the brief segments within our dataset. Subsequently, we categorize each jump in accordance with these established patterns.

To elucidate further, let us define some notation. Let s_t^e be the share of viewers experiencing emotion e at time t ; $b_t^{e,i}$ is a binary marker for the viewer i showing if they felt emotion e at time t (1 for yes, 0 for no), and n is the total viewer count for a

video. The *strength of emotion* e at a given time point t is:

$$s_t^e = \frac{\sum_{i=1}^n b_t^{e,i}}{n} \quad (1)$$

Figure 2 provides a graphical representation, illustrating the fluctuations in the strength of emotional signals over time based on the content of a specific advertisement.

Our methodology predominantly focuses on analyzing short video segments, specifically segments that are 5-seconds in length. This concise duration facilitates the capture of rapid emotional transitions, which might otherwise be diluted in longer segments. Additionally, the granularity of these segments allows us to correlate emotional shifts with specific moments in the video, shedding light on the most evocative sections. To dissect the video into these segments, we employ a sliding window approach, advancing one second at a time, thereby examining all possible 5-second combinations within a video. The difference in emotional intensity at the beginning and end of each segment represents the *emotional jump*,

$$J^e = s_{start}^e - s_{end}^e. \quad (2)$$

If J^e is positive, it means viewers felt more of that emotion by the end of the clip. Upon extracting emotional jumps from each video, our pivotal task is to identify which of these signify a *noteworthy* emotional transition. For each emotion, we analyze its corresponding jumps within the empirical frequency distributions. If a given J^e resides within the top k th quantile of its associated emotion’s distribution, it is assigned a corresponding label, highlighting a significant emotional jump.

Transitioning from continuous to discrete signals was a considered decision, driven in large part by the inherent challenges of continuous emotional metrics. These signals, despite their granularity, often contain noise from multiple sources. For one, viewers might not uniformly respond to the same video stimuli at identical times, leading to variances in emotional trajectories. Additionally, sporadic viewer reactions might not necessarily correlate with the overarching sentiment of the video. Such noise complicates subsequent analysis, obscuring genuine viewer sentiment. In contrast, discrete labels offer a distilled representation of emotions, mitigating ambiguity and creating clear boundaries between distinct emotional states. By basing our analysis on the full distribution of emotional jumps, we can more authentically capture viewer sentiment, avoiding fleeting or anomalous blips. Moreover, discrete labels align smoothly with machine learning frameworks tailored for classification tasks, simplifying both modeling and interpretation. By setting a threshold at the upper k th quantile, our methodology focuses on labeling only the most pronounced emotional shifts as positive examples. This makes the model more robust against noise and improves generalizability by ensuring it learns to recognize strongly evident emotional cues. A higher labeling threshold yields cleaner, better-defined training data consisting of salient examples, rather than attempting to fit indistinct or borderline cases.

Emotional jump classification

Given a dataset $\mathcal{D} = \{v_i, y_i\}_{i=1}^N$ of N 5-second video clips v_i , each labeled with one of 8 discrete emotions y_i , our primary goal is to predict the conditional probability distribution $p(y|v)$ for an input clip v . Formally, we want to learn a function $p = f(v)$ that outputs an 8-dimensional probability vector p given v . The complex spatio-temporal patterns in this visual classification task warrant a computer vision model capable of capturing nuanced features.

Convolutional neural networks (CNNs) are well-suited due to their hierarchical representations and exceptional ability to extract visual features, as evidenced across diverse applications⁵⁻⁷. Specifically, we employ the Temporal Shift Module (TSM)⁸ which leverages 2D CNNs with frame shifting to achieve computational efficiency without compromising representational power for video tasks like action recognition. This aligns with our insight that video-evoked emotions depend on both immediate frames and preceding context. However, in its original form, TSM lacks integration of audio, which often serves as an emotional amplifier. To address this, we enhance TSM by converting audio to mel-spectrogram representations and adding them as input frames. Mel-spectrograms capture perceptual properties of human hearing for intuitive audio features. This multimodal architecture, termed TSAM, synergistically fuses visual and auditory modalities.

In our TSAM architecture, we utilize ResNet50⁹ as the backbone to jointly process video frames v and audio mel-spectrograms a . While v undergoes temporal shifting per TSM, a bypasses this. Both modalities pass through the same ResNet50 weights, enabling shared spatio-temporal feature extraction. Post-processing dropout is applied to mitigate overfitting. We then aggregate the visual and auditory representations by averaging their features. Finally, a fully-connected layer generates predictions $\hat{y}$ for the 8 emotion classes based on this multimodal embedding. We train end-to-end by minimizing the cross-entropy loss $L(\hat{y}, y)$ between predictions $\hat{y}$ and ground truth labels y . Figure 5 presents a schematic of this architecture. The implementation of the TSAM architecture has been released as an open-source project (<https://github.com/aav-antonov/TSAM>).

For training, validation, and testing, we sample a fixed number of frames k from each input video v using uniform sampling. Specifically, we divide v temporally into k equal segments, then randomly select one frame per segment. This generates a representative subset of k frames covering the full v . Our frame dataset consists of $256 \times W$ resolution images, with variable W

based on original video aspect ratios. Keeping k consistent enables fair evaluation across diverse videos. The 256 pixel height allows efficient batch processing on GPUs. Overall, this uniform subsampling extracts informative training frames from v .

During training, we use data augmentation including random 224×224 cropping and horizontal flipping to improve robustness. For validation and testing, we use centralized 224×224 crops to standardize inputs. We optimize hyperparameters like dropout, batch size, learning rates, and epochs using the validation accuracy. The best performing model on the validation set is selected as the final model for testing. This augmentation and tuning procedure enhances diversity and reduces overfitting, resulting in optimal model generalization.

Full-length video analysis using sliding windows

Thus far, we have utilized a convolutional neural network (CNN) trained on 5-second video clips to classify emotions. However, comprehensively analyzing video advertisements requires modeling emotional responses evoked by full-length videos. Directly applying our clip-trained CNN poses challenges. Emotions manifest non-uniformly in longer videos, with certain segments eliciting intense yet transient responses while other periods remain neutral. Brief emotional peaks may become obscured when analyzing full videos holistically. Furthermore, temporal dynamics and context between segments provide important emotional insights.

To enable nuanced modeling of emotion responses to full video ads, we propose a sliding window technique. This approach synergistically combines fine-grained localized analysis of segments with consolidated scores summarizing dynamics across the full video duration. Specifically, it involves extracting overlapping 5-second segments across the video, progressing in 1-second intervals between segments. For a 30-second video, this process yields 26 overlapping windows spanning its entire length. Subsequently, each segment is classified independently by our TSAM, producing an 8-dimensional probability vector indicating the likelihood of each emotion. If an emotion's probability exceeds 0.5 for a given segment, that emotion is designated as present. After assessing all windows, we consolidate the counts C^e for each emotion and normalize by the video duration as $S^e = \frac{C^e}{T-4}$ where T is duration and $T-4$ is the window count. An emotion is deemed present if its score meets threshold t_C .

This multifaceted approach has several motivations. Video ad emotions often manifest non-uniformly, with transient yet intense bursts in certain segments. Analyzing overlapping windows helps capture even brief peaks, retaining nuance. Furthermore, we want an expansive perspective of the video's overall emotions, not just isolated segments. This consolidated view allows holistic analysis to balance detecting the overarching tone and local peaks.

Results

Descriptive statistics of viewer clicks and emotional jumps

To begin our analysis, we first examine the distribution of video lengths present in our dataset. Table 1 shows the distribution of video ads by duration, in seconds. As illustrated, our dataset is composed primarily of 30-second commercials, followed by 15-second and 60-second ads. Figure 8 depicts the distribution of number of clicks per user for each of the eight emotions, without normalizing for differences in video duration. When we normalize for a standard 30-second video length and look at the average number of clicks per user in Figure 6, happiness, surprise, and sadness emerge as the most frequently expressed emotions. Overall, our dataset provides a rich portrait of emotional responses, with happiness, surprise and sadness being the predominant emotions expressed. Although viewers could indicate "no emotion," the data reveals they did not often return to a neutral state after reporting an initial emotion. On average, users expressed 1.4 emotions per 30-second clip, while selecting "no emotion" only 0.3 times.

The empirical distribution of 5-second clips based on the percentage of users expressing various emotions is reported in Figure 7. This graph illustrates the intensity of viewers' emotional responses as a percentage alongside the proportion of video clips that triggered each level of response. Clips in the top 0.5% of the distribution, highlighted in red, were used to define the emotional jumps, which were labeled for classifier training and testing.

By applying a 5-second sliding window to all video clips, we extracted approximately 711,000 labeled segments. Table 2 provides the number of unique 5-second clips inducing an emotional shift for each emotion, along with the total number of distinct full videos containing these 5-second clips. To understand how emotional jumps appear, Figure 3 depicts instances of jumps for happiness and sadness, highlighting moments where a notable surge in a specific emotion is observed. Figure 4 presents sample frames which correspond to those emotional jumps. The purpose of showcasing these frames is to provide examples of what viewers were observing when they selected the respective emotions, offering a visual reference for the emotional peaks recorded.

Performance of emotional jump classification

For training and testing purposed, we initially split the entire set of 30,751 video ads into three sets of randomly assigned clips: training (80%), validation (10%), and testing (10%). The classification model was trained and validated on the first two sets, then tested on the held-out testing set. In this section, we report on the classification results and study their dependency

on three important design factors: the number of video frames sampled within each clip, the choice of the neural network's pretraining weights, and the modality of input used for classification, whether solely video or video in conjunction with audio. Our primary metric to assess emotion prediction performance from short clips was the classification accuracy—this metric aligns with conventional norms adopted in tasks like action recognition^{8,10–12}.

Drawing from the wealth of prior research on video-based action recognition, it's known that augmenting the frame samples and harnessing pretrained neural networks, especially with ImageNet weights¹³, generally elevates classification efficacy. This is observed in comparison to limiting the frame samples or initializing weights randomly^{8,12}. A noteworthy mention is the ResNet50 network pretrained on ImageNet21K (INET21K), which has garnered acclaim for its performance across a spectrum of benchmark computer vision tasks^{14,15}. A comprehensive view of our classification outcomes for the test set can be found in Table 3. The peak performance, a commendable 43.6%, was realized with a combination of 16 frames, both video and audio modalities, and a network trained on INET21K weights. Our analyses underscore the importance of a multimodal strategy, as harnessing both video and audio channels demonstrably uplifted the classification performance. Using multimodal inputs led to gains of around 3–4% across configurations compared to unimodal inputs. Interestingly, the benefits of increasing frame counts appeared to plateau – going from 4 to 8 frames provided minor enhancements, while further additions stagnated. Furthermore, the marginal gain from INET21K weights over ImageNet was minimal, only around 0.4%.

A deeper analysis of accuracy by emotion, shown in Table 4, reveals the classifier performed best at identifying sadness, surprise, fear, and happiness. The high performance for detecting these emotions may stem from their common occurrence and expression in the video clips. As these tend to be more universally felt and displayed compared to anger or disgust, for example, the model likely benefits from their stronger representation in the dataset. The abundance of sadness, surprise, fear, and happiness examples enables more robust learning of associated visual and audio cues. Overall, our multimodal approach leverages complementary video and audio signals to improve classification, although frame inputs exhibited diminishing returns. Finally, Supplementary Material Table 2 shows the average classification accuracy obtained on the test set when using different cutoff percentiles of the emotional jump distribution to define positive examples for model training.

Emotional jumps and *Star Rating*

The star rating of a video ad serves as a widely recognized metric for ad quality in the marketing sector. Marketing consultancy agencies often strive to achieve higher star ratings for ads, as they reflect user perceptions and are calculated by averaging users' emotional responses. System1's *Star Rating* score ranges from 1 to 5+, where 5+ signifies the highest possible quality and 1 represents the lowest. Five-star ads typically evoke happiness in the audience, leading to an increased likelihood of future commercial success. The Test Your Ad dataset annotates all videos with star ratings and emotional responses. To underscore the importance of emotion jumps, we highlight a strong correlation between the presence or absence of emotion jumps in an ad and its star rating.

Our findings are detailed in Supplementary Material Table 1, where we divided all videos from the Test Your Ad dataset into five classes based on their star ratings. The first column of the table displays the distribution of videos within each corresponding star rating class (1+ signifies a rating between 1 and 2, 2+ between 2 and 3, and so on). As observed, 307 ads achieved the highest 5+ star rating, while 15,186 ads were rated 1+. Subsequently, for each class, we calculated the proportion of videos containing at least one emotion jump associated with a specific emotion. The second column presents the respective proportions for happiness jumps. Notably, 48.8% of top-rated video ads (5+) feature at least one happiness-related emotion jump, whereas less than 1% of videos in the 1+ category exhibit a similar characteristic.

Predicting emotional jumps in full-length video ads

In previous sections, we focused on classifying short video clips for emotional jumps. Now, we extend our efforts to the challenge of predicting these jumps in full-length video ads. Recognizing emotional jumps such as those indicating happiness can serve as a gauge for ad quality and viewer engagement. In our methodology, we first detect emotional jumps per time series clip within each ad using our refined classifier. We then aggregate these detected jumps to enrich our prediction signal—for instance, by counting the total number of jumps per emotion.

Our performance assessment relies on the Area Under the Curve (AUC), a robust measure of the model's discriminative power between ads with and without specified emotions. Across the test set's full-length video ads, we observed a significant variance in the model's performance by emotion. Sadness prediction was notably accurate, with an AUC of 0.88—the highest among the emotions. Happiness and fear also yielded strong AUC scores of 0.78 and 0.83, respectively, underscoring the model's capability in detecting these emotional changes. Contrastingly, the model's performance was less impressive for other emotions: contempt and anger both registered AUC scores of 0.67, with disgust slightly above at 0.69. Surprise was identified with moderate accuracy, as reflected by an AUC of 0.73. These results imply that certain emotions, especially those that are subtler like contempt or less defined like neutral—the latter of which is implied to approach the randomness of its detection—pose greater challenges for the model. The corresponding ROC curves for all the emotions can be found in Figure 9.

By averaging the AUC scores across all emotions (excluding neutral), we obtained a mean of 0.75. This average highlights the model's overall effectiveness in recognizing emotional variations across a broad range of emotions. However, it also points to a potential area for refinement, especially in identifying emotions that are inherently more nuanced or less pronounced in video advertisements.

TSAM performance on benchmark video datasets

Despite our primary focus on emotional jumps in ads, we conducted a validation exercise by testing our TSAM model on two well-established video classification benchmarks, and specifically for action classification. These public dataset included audio, and served two key purposes. First, it enabled assessing the model's adaptability to new tasks beyond its original purpose. Second, it highlighted the value of our multimodal approach incorporating audio cues. While not directly related to identifying emotions, performance on these benchmarks was important to verify the model's performance and versatility before deployment on our primary task.

Specifically, in our experiments, we used two well-known action recognition benchmarks: Kinetics-400 and Something-Something V1. Kinetics-400 consists of 400 human action classes, while Something-Something V1 comprises 174 fine-grained actions. The large number of classes make these challenging benchmarks to evaluate model capabilities. Both focus on classifying directly observable physical motions and object interactions. In contrast, our task of emotion classification deals with inferring the abstract affective state expected to be evoked in viewers. This is substantially more challenging than categorizing visible actions and behaviors. By first validating TSAM on established action datasets, we verified model capabilities on complex tasks with many classes before specializing for nuanced emotion identification, which constitutes a more abstract, complex problem despite fewer output categories.

On Kinetics-400, Supplementary Material Table 3 shows TSAM achieves substantial accuracy gains of 1.5-2% by incorporating audio and 2-2.5% with ImageNet21K weights over ImageNet 1K. These results showcase the benefits of TSAM's multimodal design and large-scale pretraining. Additionally, Supplementary Materials Table 4 reveals TSAM attained significant performance improvements on Something-Something V1 by incrementally increasing temporal aggregation (raising k) from 1 to 4, unlike the limited k in traditional TSM. Remarkably, even without optical flow inputs, RGB-only TSAM ranks among top-performing 2D CNN architectures that utilize motion features. Our results demonstrate competitiveness at recognizing granular actions, providing an important baseline prior to emotion classification. The results also confirm the benefits of our multimodal, temporally-aware approach before specializing for emotion classification.

Discussion

Significance of the Results

Effective video advertisements rely on evoking genuine emotional reactions from viewers. Existing research^{16,17} has emphasized that videos eliciting robust emotional experiences—whether uplifting or distressing—are approximately twice as likely to be shared compared to less emotionally impactful ones. Thus, the ability to determine if a video advertisement will elicit notable emotional feedback is crucial in the world of impactful marketing. In our study, we have ventured into a comprehensive private database of 30,751 video ads, meticulously annotated for eight distinct emotions by an average of 75 viewers per ad. Our focus has been on pronounced emotional pulses within the data, termed "emotion jumps"—5-second excerpts extracted from the overarching video ads, considered by annotators to evoke the most salient emotional reactions. Our dataset consists of 26,635 clips, each labeled with one of the eight emotion tags reflecting the documented emotional responses. When confronted with this 8-class classification challenge, our model demonstrated an impressive 44% accuracy on the test dataset. Notably, specific emotions like Happiness and Sadness achieved individual accuracies of nearly 60%. The role of audio as an emotion identifier was evident in our findings. However, increasing the frame count (from 8 to 16) or utilizing pre-trained weights from INET21K^{14,15} only marginally influenced the classification output. This observation hints at potential saturation in classification accuracy, likely due to the inherent subjectivity of our dataset.

An "emotion jump" within a video ad signifies an intense emotional stir at some point. Evaluating our neural network's ability to predict the occurrence or absence of these emotion jumps within complete video ads, we employed a sliding window strategy and assessed its effectiveness using AUC metrics for individual emotions. Our results for emotions like Sadness, Happiness, and Fear were promising. Their predictive capability for identifying emotion jumps within entire video ads suggests practical implications.

Connection to Other Video Understanding Tasks

Our contributions lie within the domain of automated video understanding, a significant challenge at the intersection of computer vision and artificial intelligence. Video understanding extends traditional image recognition by introducing a temporal dimension, enabling the exploitation of motion and the intricate patterns that unfold over time. Among the multitude of tasks within this domain, human action recognition stands out, involving the identification of various human actions in video

frames^{10,18}. In comparison to our research, human action recognition focuses on explicit pattern recognition. Actions are observable, tangible, and can be directly inferred from visual content. Extensive efforts from the community have led to the annotation of numerous videos showcasing diverse human actions, resulting in the creation of high-quality, large-scale datasets that serve as benchmarks in the field^{19–22}.

Approaching the core of our study, our trajectory intersects with affective video content analysis, a niche area that aims to unravel how videos evoke emotions in humans and predict these emotional responses to dynamic visual stimuli^{2,23}. Another tangent within video understanding is facial emotion recognition, involving the detection of human faces in videos and decoding facial expressions to infer underlying emotions^{24,25}. It's essential to note that our study differs in focus. Instead of relying on direct visual cues related to video viewers, we predict emotions evoked by video content itself—a challenge compounded by the fact that emotions, unlike actions, are intangible entities, not directly displayed on video canvases but rather elicited within viewers.

Choice of Neural Network Architecture

Over the past decade, convolutional neural networks (CNNs) have found success in video understanding, particularly in action recognition^{8,11}. Mapping a video sequence to an outcome involves processing both spatial and temporal information. While 3D CNN architectures have been used for video feature extraction^{11,26}, they come with a high computational cost. High-performance 3D CNN architectures are often hindered by this computational expense¹². Various solutions have been proposed to address this issue^{11,19}. More recently, transformer-based video networks have shown superior performance on benchmark datasets compared to CNN networks^{27,28}. However, transformer-based architectures still come with substantial computational costs and tend to outperform CNNs only when a large amount of data is available²⁹.

In contrast, 2D CNNs offer computational efficiency. While 2D CNNs have been used to learn spatial features from individual frames^{30,31}, they cannot model temporal information on their own. To overcome this limitation, the Temporal Shift Module (TSM)⁸ was introduced, significantly improving performance on action recognition datasets, even when compared to 3D CNN architectures. TSM employs a ResNet50 architecture to process input frames in parallel and shifts features between temporally neighboring blocks before the convolutional layers. This feature exchange across neighboring frames greatly enhances the architecture's ability to learn temporal features.

In our work, we leverage recent advancements in this area and extend existing state-of-the-art architectures to suit our problem setting. We chose to adapt TSM for the classification of emotion jumps due to the trade-off between available data, computational efficiency, and classification performance. Furthermore, both visual and auditory stimuli contribute to a video's affective content³². Therefore, we aimed to jointly account for video and audio content when developing our convolutional neural network (CNN) architecture (see the appendix for more details).

Connections to Related Studies and Datasets

Affective video content analysis exists within a diverse landscape of datasets, but it often lacks depth and breadth when compared to other domains like action recognition, which boasts ample datasets³³. One notable dataset in this realm is the DEAP dataset³⁴, which provides insights into reactions to music videos through both physiological signals and viewer ratings. However, it is limited in scale, containing only 120 one-minute videos. Additionally, other datasets such as the WikiArt Emotions Dataset³⁵ and the Discrete LIRIS-ACCEDE dataset³⁶ take different approaches, focusing on art emotion annotation through crowdsourcing and affective video annotations using a pairwise comparison protocol, respectively.

Some research in affective video content analysis incorporates external mechanisms like facial expression imaging or physiological measurement devices to discern viewer reactions to videos^{37–39}. An example of this approach is the EEV dataset⁴⁰, which utilizes viewers' facial reactions for automatic emotion annotation in videos, albeit with the challenge of converting facial reactions into distinct emotion labels. Moreover, datasets like the EIM16 dataset, derived from the LIRIS-ACCEDE database, and the Extended COGNIMUSE dataset emphasize different aspects of emotional annotation. The EIM16 dataset, for instance, is designed for both short video excerpt emotion prediction and continuous emotion annotation for longer movie clips, while the COGNIMUSE dataset delves into the distinction between intended and expected emotions.

In comparison to these datasets, our dataset exhibits distinct features. Rooted in manual annotations, it ensures a richer and more personalized emotion spectrum. What sets our collection apart is its remarkable scale, comprising over 30,000 video ads—only the LIRIS-ACCEDE dataset comes close in magnitude. Furthermore, our dataset benefits from the consistency of cultural backgrounds among annotators, offering a unique perspective. This contrasts with datasets like LIRIS-ACCEDE, which source annotations from a global participant pool. The diversity in the emotional spectrum our dataset covers, spanning eight distinct emotions, establishes it as a robust and comprehensive resource for in-depth affective video content analysis.

Limitations of this Study

While our dataset represents an important step forward, we must acknowledge the inherent subjectivity in manual emotion annotation based on viewer recall. Labeling sentiments poses more reproducibility challenges compared to labeling relatively

objective actions. However, we can reasonably expect more consistent labeling for pronounced emotional reactions, where a large majority of viewers concur within a short span, as opposed to subtle responses.

To help mitigate subjectivity, we introduced "emotion jumps" - brief clips that trigger particularly intense responses tied to specific emotions. By focusing analysis on these pronounced spikes, we aimed to capture the most vivid and consistent reactions across viewers. We restructured the dataset into a video classification format by selecting clips that provoked heightened emotional reactions for each category. This process yielded improved consistency and reliability in the labels compared to subtle responses. It enabled our model to more effectively discern emotions evoked by videos based on detected patterns in these pronounced responses.

Our approach analyzes full-length videos by extracting overlapping 5-second segments. This sliding window technique aims to capture even brief, transient emotional peaks that non-overlapping segments would likely overlook. Focusing solely on complete videos risks missing these potent yet ephemeral reactions. However, overlapping windows introduce potential drawbacks like redundancy and overemphasis on momentary blips rather than overall sentiment. Ultimately, the sliding window methodology strikes an effective balance. It combines fine-grained localization of emotional bursts with consolidation to assess overall video sentiment. This twin perspective enables nuanced evaluation attuned to the intricacies of emotion dynamics in videos. Alternative methods may neglect critical nuances - either losing transient peaks in holistic views or lacking broader context from non-overlapping segments. Our approach fuses detailed and big picture analysis to provide the layered insight essential for video emotion understanding.

Conclusions

Our prototype represents an original advancement in emotion recognition, leveraging deep learning to predict emotional reactions directly from video content. By analyzing multimedia signals to identify moments that elicit strong responses, our model enables advertisers to create more impactful campaigns that intimately resonate with target audiences. By utilizing our evoked emotion recognition model, marketing professionals can rapidly analyze vast video databases, identifying moments that trigger powerful emotional responses. These emotionally charged segments serve as valuable resources for marketers, enabling them to create compelling and influential advertisements that resonate seamlessly with viewers. The implications are manifold—not only enhancing marketing effectiveness but also fostering deeper and more meaningful engagements with the intended audience. Beyond advertising, applications are far-reaching in luxury brand marketing, entertainment analytics, and empathetic AI systems.

There are numerous avenues for future exploration and improvement. A pressing need exists to elucidate our model's inner workings and decision-making rationale. As it currently operates as a *black box*, deciphering techniques to highlight precise emotion-evoking frames or elements within those frames would provide valuable insights. Illuminating which visual motifs or ambiance the model associates with certain emotions could allow advertisers to refine strategies for maximal emotional impact. Diversifying the cultural demographics in our dataset is another important area for improvement. While currently spanning the UK and US, expanding to broader global regions could enhance applicability. Emotional triggers and norms differ significantly across cultures. For instance, reactions in Asia or South America may deviate markedly from the current Anglo-centric data. By diversifying cultural representation, our model could learn more universal emotional patterns.

In summary, this work represents an important step forward in emotion recognition by demonstrating the ability to predict viewer sentiment directly from video content. Our deep learning model analyzes multimedia signals to uncover emotionally impactful moments that can inform more resonant advertising and content. While significant challenges remain, this pioneering research establishes a crucial foundation and agenda for future cross-disciplinary advances.

References

1. Madhumita, M. Emotion recognition: can ai detect human feelings from a face? *The Financial Times* (2021).
2. Baveye, Y., Chamaret, C., Dellandrea, E. & Chen, L. Affective video content analysis: A multidisciplinary insight. *IEEE Transactions on Affect. Comput.* **9**, DOI: [10.1109/TAFFC.2017.2661284](https://doi.org/10.1109/TAFFC.2017.2661284) (2018).
3. Ekman, P. *Emotions revealed: recognizing faces and feelings to improve communication and emotional life* (Times Books/Henry Holt and Co, 2003).
4. Posner, J., Russell, J. & Peterson, B. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev Psychopathol.* **17**, DOI: [10.1017/S0954579405050340](https://doi.org/10.1017/S0954579405050340) (2005).
5. Oprea, S. *et al.* A review on deep learning techniques for video prediction. *CoRR abs/2004.05214* (2020). [2004.05214](https://arxiv.org/abs/2004.05214).
6. Liu, W. *et al.* A survey of deep neural network architectures and their applications. *Neurocomputing* **234**, DOI: [10.1016/j.neucom.2016.12.038](https://doi.org/10.1016/j.neucom.2016.12.038) (2017).
7. Li, K. *et al.* Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations* (2022).
8. Lin, J., Gan, C. & Han, S. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, DOI: [10.1109/ICCV.2019.00718](https://doi.org/10.1109/ICCV.2019.00718) (2019).
9. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, DOI: [10.1109/CVPR.2017.634](https://doi.org/10.1109/CVPR.2017.634) (2017).
10. Singh, T. & Vishwakarma, D. K. Video benchmarks of human action datasets: a review. *Artif. Intell. Rev.* **52**, DOI: [10.1007/s10462-018-9651-1](https://doi.org/10.1007/s10462-018-9651-1) (2019).
11. Feichtenhofer, C. X3D: Expanding Architectures for Efficient Video Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, DOI: [10.1109/CVPR42600.2020.00028](https://doi.org/10.1109/CVPR42600.2020.00028) (2020).
12. Feichtenhofer, C., Fan, H., Malik, J. & He, K. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, DOI: [10.1109/ICCV.2019.00630](https://doi.org/10.1109/ICCV.2019.00630) (2019).
13. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. DOI: [10.1109/cvpr.2009.5206848](https://doi.org/10.1109/cvpr.2009.5206848) (2010).
14. Ridnik, T., Ben-Baruch, E., Noy, A. & Zelnik-Manor, L. Imagenet-21k pretraining for the masses (2021). [2104.10972](https://arxiv.org/abs/2104.10972).
15. Wightman, R. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, DOI: [10.5281/zenodo.4414861](https://doi.org/10.5281/zenodo.4414861) (2019).
16. Wood, O. *Lemon: How the Advertising Brain Turned Sour* (Institute of Practitioners in Advertising, 2019).
17. Nelson-Field, K. *The attention economy and how media works: Simple truths for marketers* (Springer Nature, 2020).
18. Zhu, Y. *et al.* A comprehensive study of deep video action recognition. *CoRR abs/2012.06567* (2020). [2012.06567](https://arxiv.org/abs/2012.06567).
19. Carreira, J. & Zisserman, A. Quo Vadis, action recognition? A new model and the kinetics dataset. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, DOI: [10.1109/CVPR.2017.502](https://doi.org/10.1109/CVPR.2017.502) (2017).
20. Goyal, R. *et al.* The 'Something Something' Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, DOI: [10.1109/ICCV.2017.622](https://doi.org/10.1109/ICCV.2017.622) (2017).
21. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. & Serre, T. HMDB: A large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, DOI: [10.1109/ICCV.2011.6126543](https://doi.org/10.1109/ICCV.2011.6126543) (2011).
22. Soomro, K., Zamir, A. R. & Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild (2012). [1212.0402](https://arxiv.org/abs/1212.0402).
23. Hanjalic, A. & Xu, L. Q. Affective video content representation and modeling. *IEEE Transactions on Multimed.* **7**, DOI: [10.1109/TMM.2004.840618](https://doi.org/10.1109/TMM.2004.840618) (2005).
24. Mellouk, W. & Handouzi, W. Facial emotion recognition using deep learning: Review and insights. In *Procedia Computer Science*, vol. 175, DOI: [10.1016/j.procs.2020.07.101](https://doi.org/10.1016/j.procs.2020.07.101) (2020).
25. Bolcas, R. D. & Dranga, D. Facial Emotions Recognition in Machine Learning. *EEA - Electrotehnica, Electron. Autom.* **69**, DOI: [10.46904/eea.21.69.4.1108010](https://doi.org/10.46904/eea.21.69.4.1108010) (2021).

26. Rahmani, H. & Mian, A. 3D action recognition from novel viewpoints. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, DOI: [10.1109/CVPR.2016.167](https://doi.org/10.1109/CVPR.2016.167) (2016).
27. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR* **abs/2010.11929** (2020). [2010.11929](https://arxiv.org/abs/2010.11929).
28. Li, K. *et al.* Uniformer: Unified transformer for efficient spatiotemporal representation learning. *CoRR* **abs/2201.04676** (2022). [2201.04676](https://arxiv.org/abs/2201.04676).
29. Han, K. *et al.* A survey on vision transformer. *IEEE Transactions on Pattern Analysis Mach. Intell.* 1–1, DOI: [10.1109/tpami.2022.3152247](https://doi.org/10.1109/tpami.2022.3152247) (2022).
30. Wang, L. *et al.* Temporal segment networks: Towards good practices for deep action recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9912 LNCS, DOI: [10.1007/978-3-319-46484-8_12](https://doi.org/10.1007/978-3-319-46484-8_12) (2016).
31. Wang, L. *et al.* Temporal Segment Networks for Action Recognition in Videos. *IEEE Transactions on Pattern Analysis Mach. Intell.* **41**, DOI: [10.1109/TPAMI.2018.2868668](https://doi.org/10.1109/TPAMI.2018.2868668) (2019).
32. Naz, H. & Ahuja, S. Latest trends in emotion recognition methods: case study on emotiw challenge. *Int. J. Adv. Comput. Res.* **10**, DOI: [10.19101/ijacr.2019.940117](https://doi.org/10.19101/ijacr.2019.940117) (2020).
33. Zhao, S., Wang, S., Soleymani, M., Joshi, D. & Ji, Q. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Transactions on Multimed. Comput. Commun. Appl.* **15**, 1–32, DOI: [10.1145/3363560](https://doi.org/10.1145/3363560) (2019).
34. Koelstra, S. *et al.* DEAP: A database for emotion analysis; Using physiological signals. *IEEE Transactions on Affect. Comput.* **3**, DOI: [10.1109/T-AFFC.2011.15](https://doi.org/10.1109/T-AFFC.2011.15) (2012).
35. Mohammad, S. M. & Kiritchenko, S. Wikiart emotions: An annotated dataset of emotions evoked by ART. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation* (2019).
36. Baveye, Y., Dellandréa, E., Chamaret, C. & Chen, L. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affect. Comput.* **6**, DOI: [10.1109/TAFFC.2015.2396531](https://doi.org/10.1109/TAFFC.2015.2396531) (2015).
37. Sartori, A. *et al.* Looking at mondrian's victory boogie-woogie: What do i feel? In *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2015-January (2015).
38. Jiang, Y. G., Xu, B. & Xue, X. Predicting emotions in user-generated videos. In *Proceedings of the National Conference on Artificial Intelligence*, vol. 1 (2014).
39. Schaefer, A., Nils, F., Philippot, P. & Sanchez, X. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cogn. Emot.* **24**, DOI: [10.1080/02699930903274322](https://doi.org/10.1080/02699930903274322) (2010).
40. Sun, J. J. *et al.* EEV Dataset: Predicting Expressions Evoked by Diverse Videos. *arXiv* (2020).

Author contributions statement

A.A. spearheaded the model implementation, conducted comprehensive data analysis, and authored the initial manuscript draft. S.S.K., W.H., and O.W. played pivotal roles in facilitating data access, providing valuable insights into data interpretation, and actively collaborated by offering comments on the manuscript. G.M., in addition to securing funding, offered critical oversight and expert guidance in all aspects of data modeling for the study, and actively collaborated in manuscript composition.

Additional information

Competing Interests Statement: S.S.K., W.H., and O.W. are employees of the company that provided the dataset for this study. Their roles were primarily focused on facilitating data access and providing input on data interpretation but did not involve direct involvement in the development of the models or the influence of the study's results.

Data availability

The data analyzed in this study is proprietary and was obtained from System1 Research under license for the current research. Restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. However, the data are available from the authors upon reasonable request and with permission of System1 Research by contacting: legal@system1group.com.

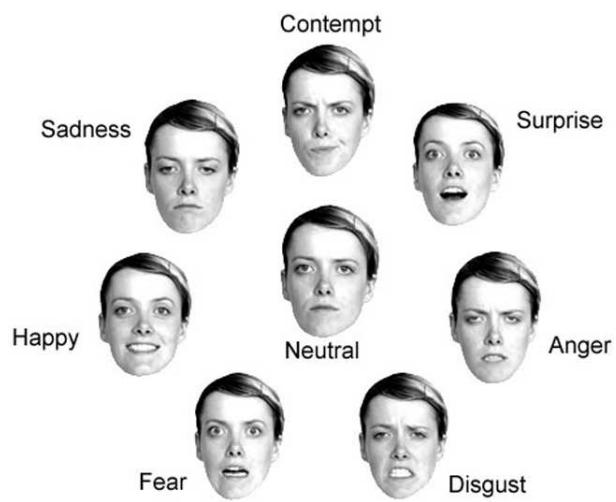


Figure 1. Facial expressions used by System1 Research's FaceTrace® method during the video annotation process. (Source: System1 Research ® 2006; reproduced with permission)

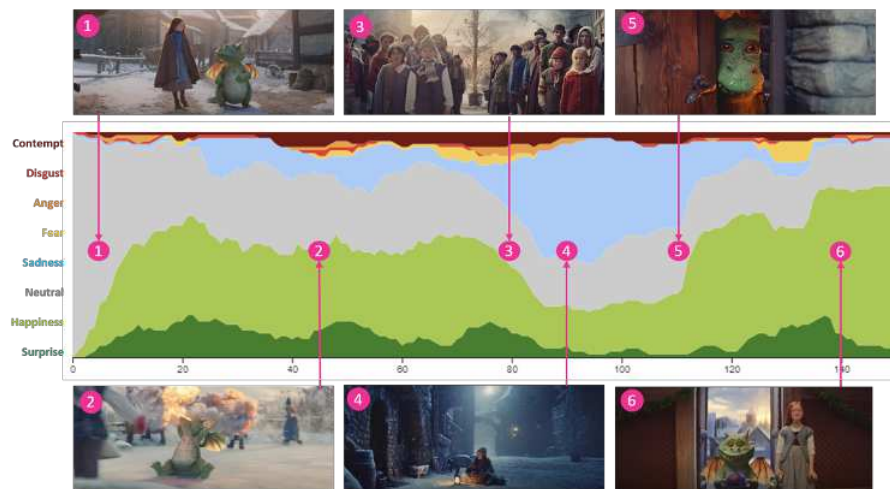


Figure 2. Test Your Ad data: At each time point throughout a video clip, we can measure the proportion of viewers in the panel (approximately $n = 75$) who self-declared experiencing one of the eight emotions. This example illustrates the changes in emotional profiles within a video.

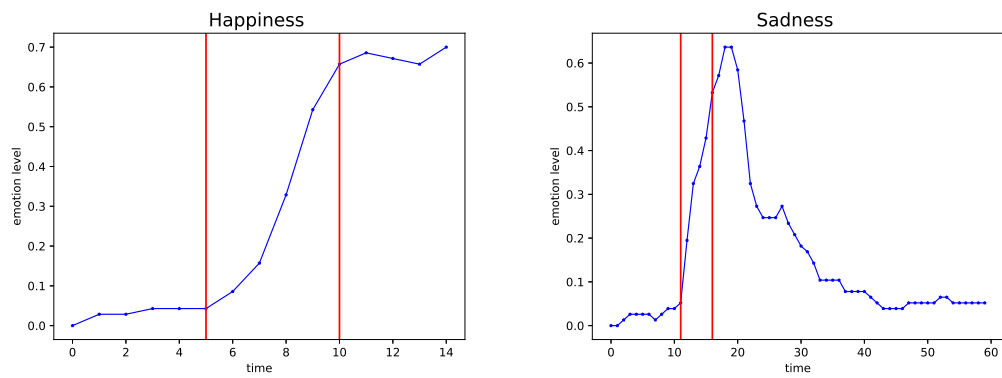


Figure 3. Test Your Ad data: exemplar emotional jumps for happiness and sadness. An emotional jump is defined as a substantial surge in the proportion of viewers who selected a particular emotion within a discrete 5-second interval.



Figure 4. Test Your Ad data: exemplar frames associated with emotional jumps for happiness (left) and sadness (right). The corresponding jumps are those illustrated in Figure 3.

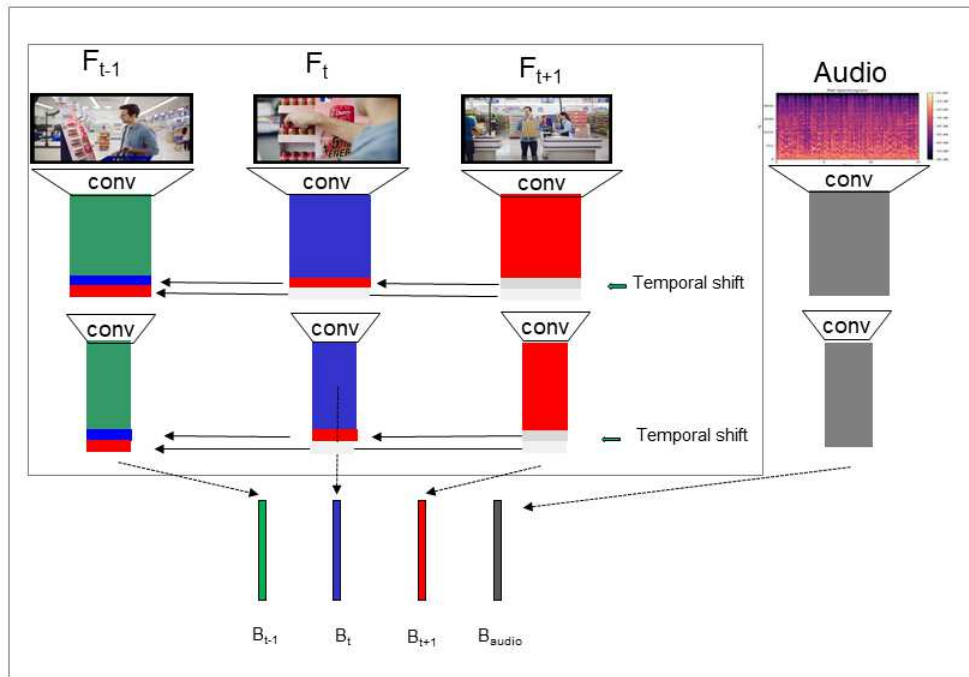


Figure 5. TSAM model: the multi-modal CNN architecture takes as input a predefined number of video frames (video segments) and audio converted into mel-spectrograms (audio segments). The ResNet50 backbone is used to extract features from both video and audio segments. Features from video segments are shifted between each other at different blocks of ResNet50. The audio input is represented by the mel-spectrogram and is processed by the same backbone without shifting. The extracted features are fused by averaging and mapped to the output classes using a fully connected layer.

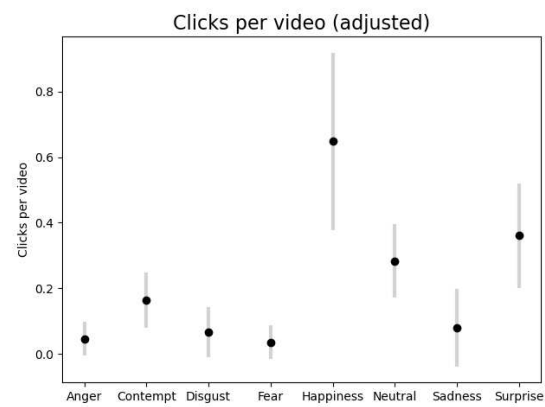


Figure 6. Test Your Ad data: average number of user clicks per emotion for every 30 seconds of video, adjusted to account for different video lengths.

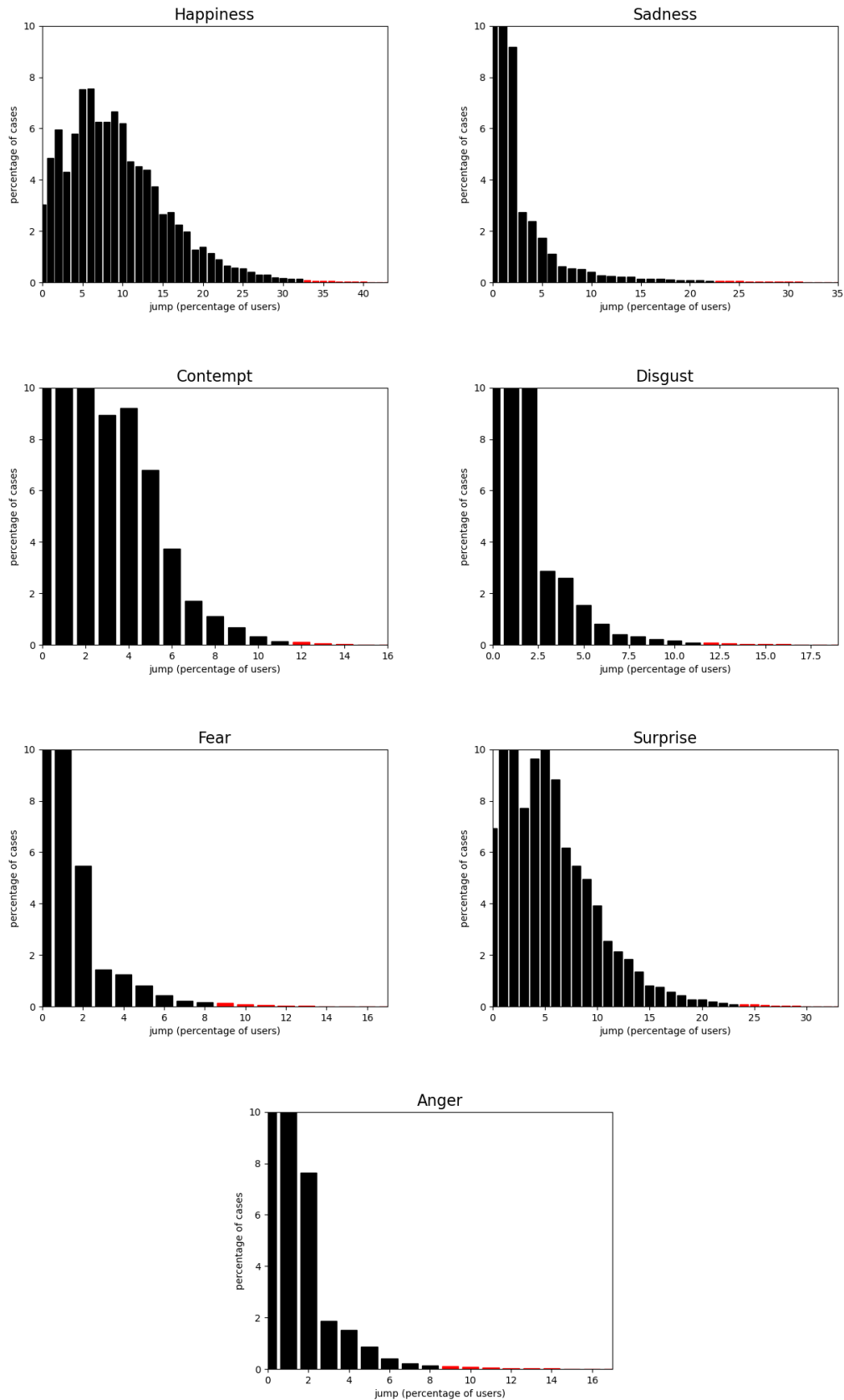


Figure 7. Test Your Ad dataset: Distribution of 5-second clips based on the percentage of users expressing various emotions. The x-axis shows the response strength as the percentage of viewers feeling the emotion. The y-axis shows the percentage of clips evoking that response level. Clips in the top 0.5% of the distribution (highlighted red) were used to define the emotional jumps, which were labeled for classifier training and testing.

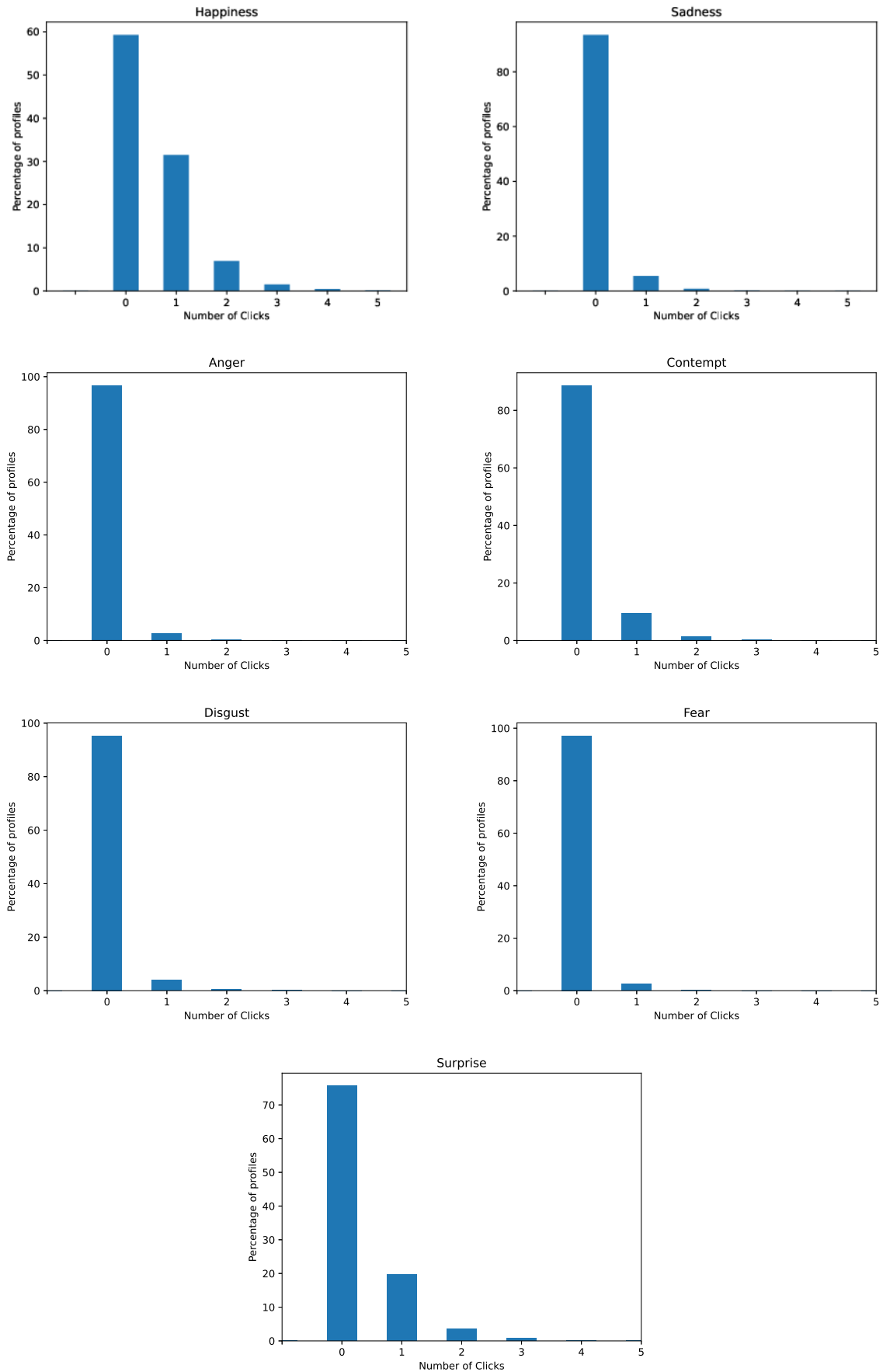


Figure 8. Test Your Ad data: distribution of number of emotion clicks per user per video, not adjusted for video duration. 10/11

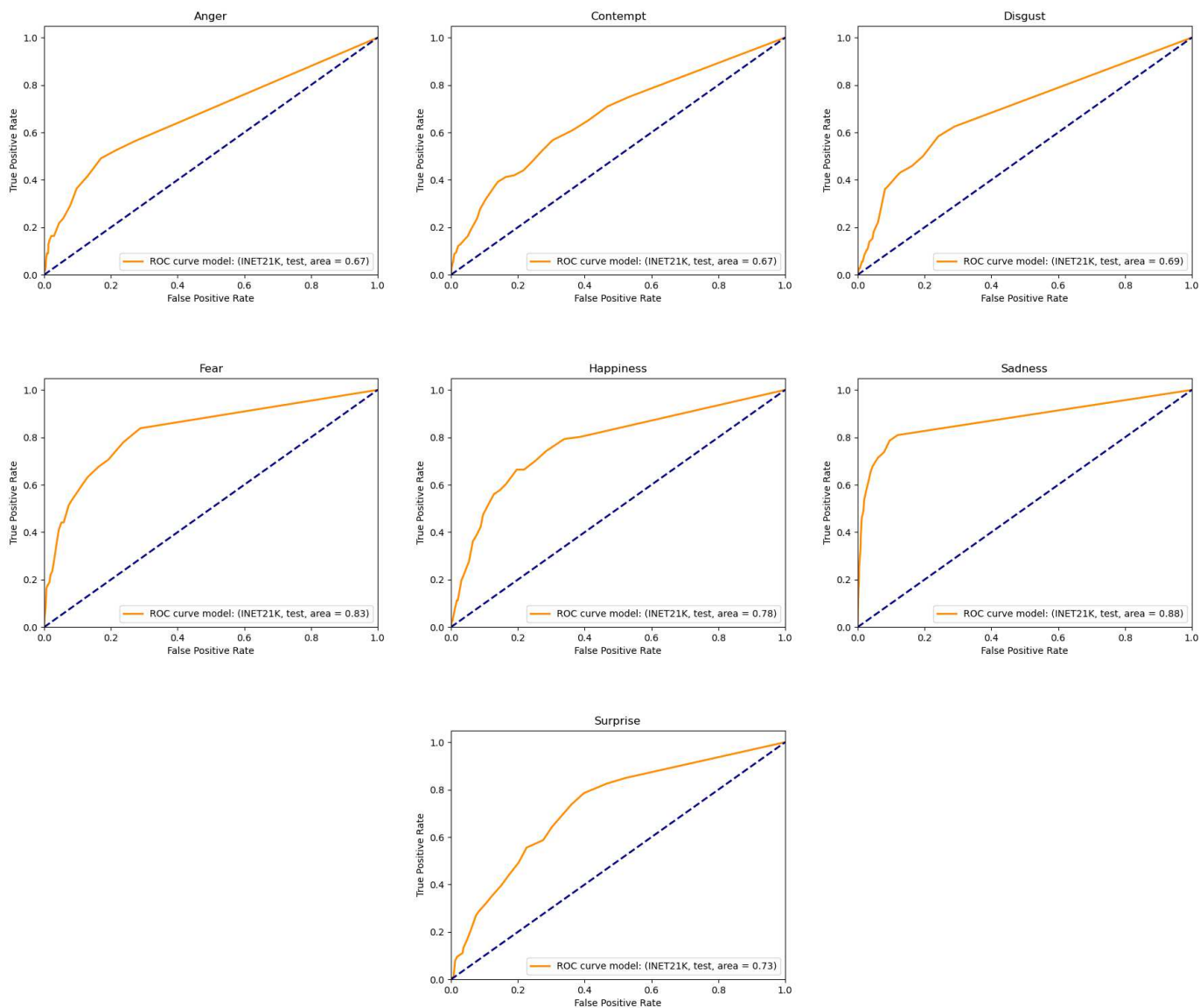


Figure 9. ROC curves for predicting the presence of emotion jumps in full-length video ads using our best CNN model (16 frames, RGB + audio input, pretrained on INET21K).

Duration	Number of Videos
5	229
10	1,640
15	8,548
20	1,811
30	15,123
40	263
45	105
60	2,218
75	39
120	297
Other	481

Table 1. Test Your Ad data: distribution of video ads by duration (in seconds)

Emotion	No of clips	No of videos
Anger	2894	653
Contempt	3317	1385
Disgust	3061	828
Fear	3166	787
Happiness	3577	1488
Neutral	3491	1395
Sadness	3576	859
Surprise	3553	1330
Total	26635	6920

Table 2. Test Your Ad data: number of labeled 5-second video clips and corresponding number of videos, categorized by emotion

Modality	Frames	Pre-training	Accuracy
RGB	4	Imagenet	38.8
RGB+audio	4	Imagenet	42.4
RGB	8	Imagenet	38.9
RGB+audio	8	Imagenet	43.2
RGB	16	Imagenet	38.8
RGB+audio	16	Imagenet	42.2
RGB	4	INET21K	39.7
RGB+audio	4	INET21K	43.4
RGB	8	INET21K	41.2
RGB+audio	8	INET21K	43.5
RGB	16	INET21K	40.9
RGB+audio	16	INET21K	43.6

Table 3. Test Your Ad data: average accuracy of the video classifier broken down by input modality, number of video frames and pre-training method

Emotion	Balanced Accuracy	Test Set Size (clips)
Anger	28.4	208
Contempt	42.3	269
Disgust	26.7	243
Fear	50.7	300
Happiness	55.8	326
Sadness	60.2	344
Surprise	51.1	325

Table 4. Test Your Ad data: balanced accuracy and test set size achieved by the video classifier broken down by emotion.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [suppmaterial.pdf](#)