



Chapitre d'actes

2017

Published version

Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

Large-scale Affective Content Analysis: Combining Media Content Features and Facial Reactions

McDuff, Daniel; Soleymani, Mohammad

How to cite

MCDUFF, Daniel, SOLEYMANI, Mohammad. Large-scale Affective Content Analysis: Combining Media Content Features and Facial Reactions. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). Washington (DC, USA). [s.l.] : IEEE, 2017. p. 339–345. doi: 10.1109/FG.2017.49

This publication URL: <https://archive-ouverte.unige.ch/unige:98499>

Publication DOI: [10.1109/FG.2017.49](https://doi.org/10.1109/FG.2017.49)

Large-scale Affective Content Analysis: Combining Media Content Features and Facial Reactions

Daniel McDuff^{1*} and Mohammad Soleymani^{2†}

¹ Microsoft Research, Redmond, USA

² Swiss Center for Affective Sciences, University of Geneva, Switzerland

Abstract— We present a novel a multimodal fusion model for affective content analysis, combining combining visual, audio and deep visual-sentiment descriptors from the media content with automated facial action measurements from naturalistic responses to the media. We collected a dataset of 48,867 facial responses to 384 media clips and extracted a rich feature set from the facial responses and media content. The stimulus videos were validated to be informative, inspiring, persuasive, sentimental or amusing. By combining the features we were able to obtain a classification accuracy of 63% (weighted F1-score: 0.62) for a five class task. This was a significant improvement over using the media content features alone. By analyzing the feature sets independently we found that states of informed and persuaded were difficult to differentiate from facial responses alone due to the presence of similar sets of action units in each state (AU 2 occurring frequently in both cases). Facial actions (smiles in particular) were beneficial in differentiating between amused and informed states whereas media content features alone performed less well due to similarities in the visual and audio make up of the content. We highlight examples of content and reactions from each class. This is the first affective content analysis based on reactions of 10,000s of people.

I. INTRODUCTION

Three hundred hours of content is uploaded to YouTube every hour. Much of this media is consumed for entertainment or information purposes. Affective content analysis has the potential to be very beneficial for enhancing video indexing utility and search efficiency. A number of approaches for automated detection of affect from content or responses to content have been presented in recent years [1], [2], [3].

Wang and Ji [4] present a survey of affective content analysis methods from video. They highlight three methods for analysis using: 1) visual and audio features of video (stimulus) content, 2) self-reported emotional descriptors (self-report labels) of video content, 3) quantified non-verbal responses (e.g. facial expressions, physiological responses) to content. The different methods can be divided into two classes: implicit tagging, using passively observed information from the media and/or responses to it (methods 1 and 3), or explicit tagging (asking users to subjectively evaluate the affective content) [5] (method 2). Purely automated methods have the distinct advantage of not requiring laborious manual annotation. Furthermore, if only ubiquitous hardware is required (e.g., webcams) they can be highly scalable.

Typically, affective video content analyses, such as this, have involved measurement and modeling of non-verbal

*damcduff@microsoft.com. This work was completed while at Affectiva.
†mohammad.soleymani@unige.ch. The work of Soleymani is supported by his Swiss National Science Foundation Ambizione grant.

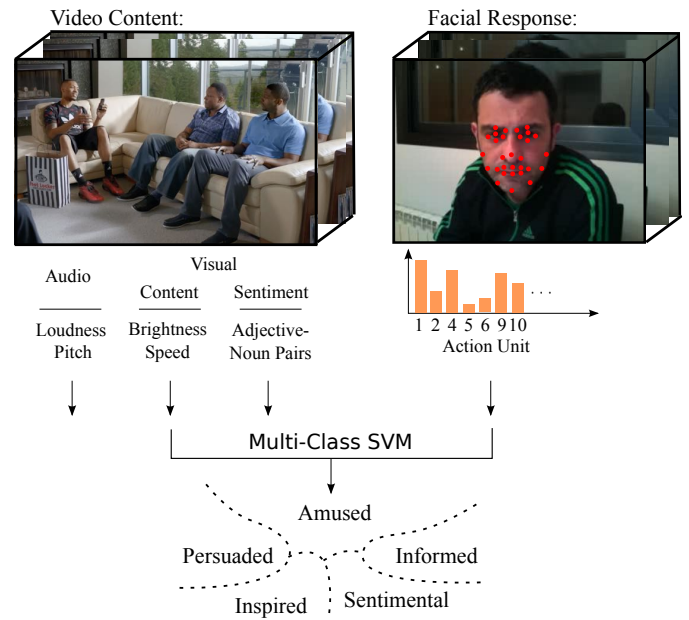


Fig. 1. Large-scale implicit affective content tagging using facial responses to media collected over the Internet. Audio and video features are extracted from the media content. Automated facial coding is used to quantify responses to the media content. Features are extracted and a classifier trained for discriminating between different types of affective content.

responses from small populations watching a limited number of video clips. This is in part due to the poor scalability of traditional methods for quantifying non-verbal signals. In Wang and Ji's survey [4] of implicit video affective content analysis research all but one study of behavioral or physiological responses to video included fewer than 100 subjects. The remaining example [6] used a similar framework to the one we leveraged in this study for crowdsourcing data. Frameworks that use the Internet and automated facial coding technology give access to large-scale observational data about viewers' responses to content [7]. This has applications for many areas of research beyond video indexing, including media measurement [8], highlight or recommendation systems [9] and the study of psychology [10].

Many studies have analyzed affective responses to emotion eliciting clips. However, due to the time consuming nature of manually coding media stimuli little has been done to jointly model aspects of media content and physiological/behavioral responses to predict affect. Soleymani *et al.* [5] presented a method for implicit categorization of movie scenes in which affective response measurements were combined with stimu-

lus video analysis for classification. We collected thousands of responses to hundreds of videos in order to explore the combination of modalities. We believe this is the largest dataset of its kind.

Over the past 50 years a majority of research into facial expressions has focused on a small set of so called “basic” emotions (joy, sadness, fear, disgust, anger and surprise). There remains disagreement over whether these states are universally expressed [11], [12], [13]; however, researchers agree that these and other expressions convey consistent information within specific contexts and that “non-basic” states warrant further study [14]. The most relevant emotional states vary by situation (e.g. when viewers watch TV ads expressions of amusement and sentimentality will be more likely than expressions of fear), except perhaps in the case of political ads. However, there is little scientific evidence for the appearance of a sentimental facial expression.

The Facial Action Coding System (FACS) [15] provides an objective and comprehensive taxonomy of actions that can be coded observationally. Action Units (AUs) are independent of emotions and therefore allow us to capture expressions of emotion that might not be part of the “basic” set. Manual FACS coding is time-consuming and laborious. However, recent advancements in computer vision have led to the development of automated facial action coding software that can detect naturalistic actions in unconstrained settings [16], [17]. As part of our analysis we examine which facial actions are predictive of the states of amused, informed, inspired, persuaded and sentimental.

The aim of this work is to evaluate a multimodal fusion model, combining contextual and facial response features, at scale on an ecologically valid dataset. Figure 1 shows an overview of our approach. We analyzed 48,867 responses to 384 video clips. The contributions of this work are: 1) presentation of a large naturalistic dataset of facial responses to online videos combined with visual, audio and deep visual sentiment descriptions of the videos, 2) design and evaluation of classifiers for prediction of affective content based on facial responses and media content descriptions, 3) a qualitative analysis of stimuli and responses for each affect class. The data (facial action measurements and media content descriptions) presented in this analysis are available at: alumni.media.mit.edu/~djmcduff/fg17.

II. RELATED WORK

Affective video content analysis entails predicting the affective response elicited in viewers by the content. The earlier work on this topic, such as [1], involved using hand-crafted content features motivated by film studies, for example, darker scenes are more likely to elicit negative emotions. More recent work on emotional content analysis in videos use deep learning and in particular convolutional neural networks which learn lower level representations automatically [18]. Despite its advantages, content analysis requires a large number of explicit ratings to be effective. Moreover, the inter-rater agreement on emotional scores is

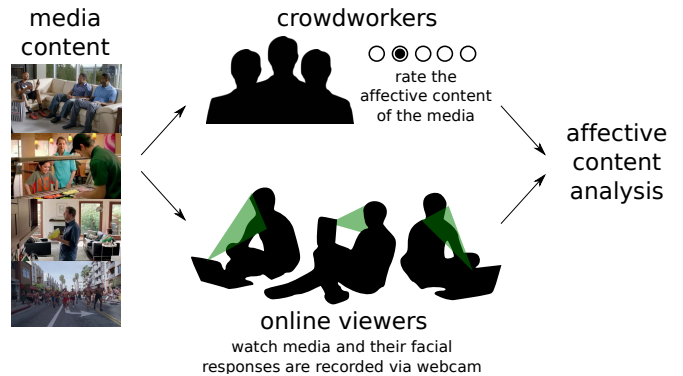


Fig. 2. For our affective content analysis, labels were collected from crowdworkers who rated the affective content of the video media. Independently viewers watched the media online and their facial responses were recorded and coded.

not always very high [19] which reduces the performance of such systems.

Implicit tagging was proposed to reduce the burden of collecting reliable explicit ratings [20]. Since implicit tagging relies on users’ natural reactions the process does interrupt their experience. Implicit tagging has been applied to different problems from emotional tagging and preference detection to topic relevance assessment [5]. Given a set of videos with known emotional tags, the association between viewers’ spontaneous reactions (e.g. facial expressions) and the emotional tags can be learned. Emotional characterization of video clips can be also performed continuously as demonstrated in [21]. Continuous characterization of videos can be used for predicting the emotional highlights in videos. Spontaneous reactions can be also used to detect action tendencies and attitudes towards the content, such as movie ratings [22] and ad effectiveness [6]. For a detailed review on emotional analysis in videos, we refer the reader to [19], [4], [5]. Emotional tagging of videos enables a multitude of applications, including: video summarization [9], affective indexing [19], and movie rating prediction [22].

III. DATA

A. Stimuli

A set of 2,215 video ads were selected as candidate stimuli that induce a range of different emotions. These videos were labeled using Amazon’s Mechanical Turk platform in order to validate the feelings that they induced. Workers watched each ad and were asked the following question:

Q. How did you feel while watching the ad (choose up to three words)?

Amused, Shocked, Persuaded, Fascinated, Inspired, Informed, Disgusted, Sentimental, Upset, Annoyed, Bored

The coders were allowed to pick up to three codes and were paid \$0.30 for coding each ad. Overall the Fleiss’ κ for the coding was moderate at 0.63. Many of the videos received different codes from different coders and in these

cases there was not 100% agreement on any one code. For a subset of 384 videos there was agreement from all the coders that one specific code applied. We used this set of videos as reliable stimuli for states of: informed (n=159), inspired (n=11), persuaded (n=37), sentimental (n=12) and amused (n=165). A thorough description of the labeling process can be found in [23].

B. Method

We recorded facial responses to the 384 video stimuli using an Internet-based framework [24]. The participants were contacted via email. Our browser interface requested use of their webcam to capture their facial response during the video. Automated analysis of naturalistic facial responses collected under unconstrained conditions presents challenges [24]. However, as our results in Section IV-B show, our system performs well on similar videos. Further information about the data collection can be found in [24].

In total we collected 48,867 facial video responses to the 384 videos. The number of facial videos for each affective state was: 19,678 (amused), 20,940 (informed), 1,771 (inspired), 4,949 (persuaded) and 1,529 (sentimental). The mean number of facial responses to each ad was 126 (std: 62). All participants were English speaking people in the United States of America. The participants represented a broad range of age groups from 16 to 65+. We made sure to balance gender and age groups but no other criteria were used.

IV. FEATURE EXTRACTION

A. Media Content

Audio, visual and mid-level sentiment visual descriptors were extracted from the ads. Table I provides a list and description of the features extracted from the media content.

1) *Audio Descriptors*: Audio descriptors were extracted from the stimuli. These capture qualities such as loudness and the power in different frequency bands. We extracted the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [25] using openSMILE [26]. These 89 features are hand-picked by experts based on their pertinence to the task of emotion recognition in speech and music.

2) *Visual Descriptors*: A set of 51 simple low-level visual descriptors were extracted from the video channel of the stimuli. These capture qualities such as motion, contrast and entropy. Visual descriptors (with the exception of the motion component) were extracted from the key frames (I-frames). For each video, visual features were pooled by averaging and calculating the standard deviation.

3) *Sentiment Content Descriptors*: A sentiment concept detector [29] was applied to key frames (I-frames) in the media content. This sentiment detector was based on a large-scale multi-lingual visual sentiment concept ontology (MVSO) that assigned adjective-noun pairs to images. Adjective-noun pairs were extracted from social media data based on their significance in expressing sentiment. The model that we used from the original work is a deep convolutional neural network (CNN). This CNN generates probability estimates for 4,342 adjective-noun pairs. We

TABLE I
STIMULI VISUAL FEATURES, THEIR DESCRIPTIONS AND DIMENSIONALITY.

Feature	Description	#
Visual features [27]	Entropy, exposure, balance, brightness, compression quality [28], contrast, sharpness, uniformity, image asymmetry (intensity), image asymmetry (histogram of gradients (HOG)), motion component (norm of difference between consecutive frames), color histogram (four bins for each color (RGB) channel), Contrast balance (Euclidean distance between the original image and the contrast-equalized image), video length, number of pixels	51
Sentiment descriptors [29]	Probabilities of adjective noun pairs related to sentiment	4342
Acoustic features	eGeMAPS[25] feature-set including pitch, loudness and Mel Frequency Cepstral Coefficients (MFCC)	89

applied the model on our data and the probability estimates were averaged for each video to form a feature vector.

B. Facial Actions and Expressions

We used automated software to code the facial actions of the viewers (Affdex - Affectiva, Inc.). Face detection was performed using the Viola-Jones method [30]. Thirty-four facial landmarks are detected using a supervised descent based landmark detector applied to the cropped face region. A refined image region of interest (ROI) was segmented using the facial landmarks. The ROI included the eyes, eyebrows, nose and mouth. The ROI was normalized using rotation and scaling to 96x96 pixels. In order to capture textural changes of the face histograms of oriented gradient (HOG) features were extracted from the image ROI. The HOG features were extracted from 32 x 32 pixel blocks (cell-size 8 x 8 pixels) with a stride of 16 pixels. A histogram with 6 bins was used for each block. This resulted in a feature vector of length 2,400 (25*16*6).




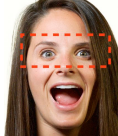













Finally, support vector machine (SVM) classifiers were used to detect the presence of each facial action (details of how the classifiers were trained and validated can be found in [16]). For each facial action a baseline appearance was estimated using a rolling 30-second window in order to account for differences in neutral appearance. The facial action classifiers returned a confidence score from 0 to 100. The software provided scores for 18 facial actions (see Table II) including 17 symmetric FACS actions (AUs 1,2,4,5,6,9,10,12,14,15,17,18,24,25,26,28,43) plus asymmetric lip corner pulls. The software is available through the AFFDEX SDK [31].

V. MODELING

The mean facial coding features and media content features were calculated for each piece of media content. This process resulted in 18 action unit values and 4,482 media content features. These are averaged across time to form a final vector for each media video.

TABLE II

DEFINITIONS AND ILLUSTRATIVE, POSED EXAMPLES OF THE FACIAL ACTIONS. ROC = AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE. TPR = TRUE POSITIVE RATE AT CLASSIFIER OPERATING POINT FOR FALSE POSITIVE RATE = .02.

	AU1	AU2	AU4	AU5	AU6	AU9	AU10	AU12	AU14
ROC/TPR	In. Brow R. .92/.62	Out. Brow R. .92/.52	Brow Lower .93/.33	Eye Widen .87/.53	Cheek Raise .72/.42	Nose Wrinke .91/.56	Up. Lip Rai. .91/.49	Lip Corner Pull .75/.36	Dimpler .89/.62
									
									Asy. AU12 Smirk .85/.44
ROC/TPR	Lip Depress. .76/.21	Chin Raise .84/.58	Lip Pucker .75/.26	Lip Press .85/.56	Lips Part .79/.41	Jaw Drop .90/.56	Lip Suck .82/.21	Eye Closure .96/.58	

Since the dimensionality of the media content features was very large relative to the total number of media units (384) we concatenated all the features and then used principal component analysis (PCA) to reduce the dimensionality to the 18 principal components with the greatest energy.

We performed classification using the facial action features, audio descriptors, video descriptors, deep visual sentiment descriptors and a combination of the features. We used a 10-fold testing scheme, holding out a random sample of 10% of the data in each repetition for testing. In the training stages the synthetic minority over-sampling technique [32] was used to over-sample examples and balance the training set class sizes. For classification we used an SVM with radial basis function (RBF) kernel. During validation (performed via a separate 10-fold validation on the training data) we optimized the cost, C , and gamma, γ , parameters.

VI. RESULTS

Due to the unbalanced class sizes (we did not over-sample the testing set) we use several performance metrics for evaluating the classification performance including confusion matrices, accuracy, and weighted and unweighted F1-scores. See Figure 5 for examples (media frames and responses) from cases that were correctly classified for each class.

A. Media Content

1) *Audio Descriptors*: Using the audio descriptors the prediction accuracy, weighted and unweighted F1-scores were 53%, 0.52 and 0.43 respectively. Figure 3(i) shows the confusion matrix.

2) *Visual Descriptors*: Using the visual descriptors the prediction accuracy, weighted and unweighted F1-scores were 54%, 0.53 and 0.43 respectively. Figure 3(ii) shows the confusion matrix.

3) *Visual-Sentiment Descriptors*: Using the deep visual-sentiment descriptors the prediction accuracy, weighted and unweighted F1-scores were 51%, 0.51 and 0.42 respectively. Figure 3(iii) shows the confusion matrix.

		Predicted					Predicted				
		Informed	Inspired	Persuaded	Sentimental	Amused	Informed	Inspired	Persuaded	Sentimental	Amused
Actual	i) Amused	110	43	3	18	3	107	42	3	21	4
	Informed	49	75	2	20	3	53	73	2	15	6
	Inspired	2	3	7	1	1	2	4	7	1	0
	Persuaded	9	19	0	12	0	11	15	0	14	0
	Sentimental	4	3	1	0	2	6	2	0	0	2
Actual	iii) Amused	105	49	5	15	3	142	12	3	8	12
	Informed	55	72	1	16	5	23	70	13	37	6
	Inspired	2	4	6	1	1	6	4	2	1	1
	Persuaded	13	16	0	11	0	8	17	4	11	0
	Sentimental	5	2	0	0	3	5	4	0	0	1
Actual	v) Amused	128	30	2	11	6					
	Informed	35	91	0	19	4					
	Inspired	3	3	8	0	0					
	Persuaded	13	16	0	11	0					
	Sentimental	6	1	1	0	2					

Fig. 3. Confusion matrices for affect classification. Using: i) media video descriptors, ii) media audio descriptors, iii) media sentiment descriptors, iv) facial action features, and v) all media descriptors and facial action features.

B. Facial Expressions

Using the facial action features the prediction accuracy, weighted and unweighted F1-scores were 59%, 0.53 and 0.43 respectively. Figure 3(iv) shows the confusion matrix. The difference between the accuracy using the facial action features and using the media content features was only significant when compared with the visual-sentiment descriptors alone ($t(9)=2.74$, $p=0.02$).

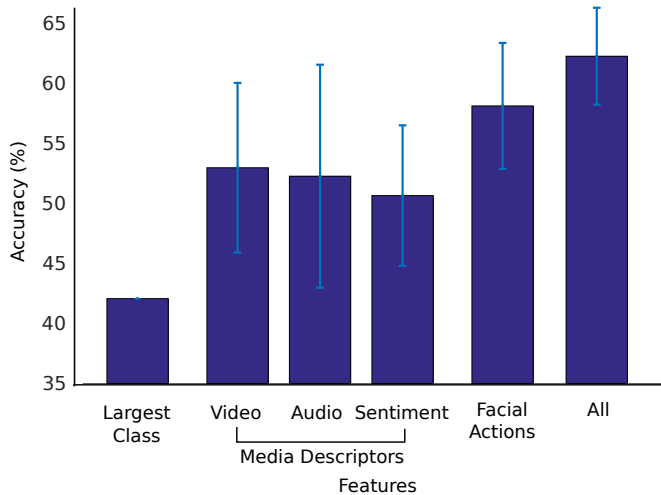


Fig. 4. Average accuracy across 10-fold testing. The error bars represent one standard deviation either side of the mean.

C. Multimodal

Using a combination of the media content and facial action features the prediction the accuracy, weighted and unweighted F1-scores were 63%, 0.62 and 0.49 respectively. Figure 3(v) shows the confusion matrix. Combining the facial and stimulus features helped correctly classify more of the media content especially for the underrepresented classes (inspired, persuaded and sentimental). Figure 4 shows the accuracy across the 10-fold testing using each set of features. The difference between the accuracy using all features and using the media content features was significant in all cases (audio, $t(9)=3.15$, $p=0.01$; video, $t(9)=4.30$, $p<0.01$; sentiment, $t(9)=5.88$, $p\ll 0.01$).

VII. DISCUSSION

Overall, the amused and informed classes were classified correctly most frequently, 72% and 61% accuracy respectively for the multimodal case. There were a greater number of examples from these classes in the training set (before oversampling). The inspired class was the next most consistently classified with 57% accuracy. The persuaded and sentimental states were not often correctly classified, 28% and 20% accuracy respectively for the multimodal case.

Using only the facial action features the underrepresented classes were frequently misclassified. The informed and persuaded classes were more often confused, as were the inspired and informed classes, suggesting that facial responses for these affective states are similar and difficult to distinguish without context. Including the content features reduced the number of informed responses misclassified as persuaded from 37 to 19 (49% reduction). Examples of common facial actions from the informed and persuaded media classes are shown in Figure 5, outer eyebrow raises (AU 2) were often present in these responses, one of the reasons they may have been confused without context.

Using the media content features the amused and informed classes were misclassified at a rate close to 35%. This suggests that visual and audio content is similar in the two

types of ads. However, introducing facial action features reduced this error rate to 23%. The rate of smiling in viewers was naturally a helpful indicator for discriminating between the amused and informed classes.

Overall, including the facial action features increased the number of correctly classified examples across all classes except the persuaded class. This suggests that for most of the classes there are consistencies in facial responses.

The sentimental class was often misclassified as amused regardless of the features used. Perhaps because smiling frequently occurs during sentimental content [23].

Whilst performance improved when features from additional modalities were included the classification is still not perfect. There may be a number of possible reasons for this. Facial responses in these states do not have a uni-modal or universal expression. There are differences between aesthetic and felt emotion. The labels in this study were perceived intended emotion; however, the felt emotion of an individual may differ from what is perceived to be the intended emotion.

VIII. CONCLUSIONS

We present a novel multimodal model for affective video content analysis using deep visual-sentiment descriptions and automated facial coding. A dataset of almost 50,000 facial responses to 384 video clips was collected using an Internet framework. Visual and audio features were extracted from the media content as were deep visual-sentiment descriptions. Automated facial coding algorithms were used to quantify 18 facial actions in each of the 23.4 million frames of facial responses that we collected to the media content. The stimulus videos were validated as informative, inspiring, persuasive, sentimental or amusing.

We built a classifier based on the automatically detected media content and facial action features to predict the affective content of each video clip. Using context from the video and audio content, in addition to information from the facial responses, led to the best performance, yielding a 63% accuracy and weighted F1-score of 0.62.

Distinguishing between affect classes of informed and persuaded using facial responses alone was particularly difficult. The typical facial responses for these classes of media content were similar, with eyebrow raising (AU 2) relatively common in both. Smiles helped discriminate between the amused and informed classes of content and facial action features improved the classification over the use of media content features alone.

Ads are a specific type of short video vignette that are designed to be likable and persuasive. Furthermore they feature certain categories of visual content (e.g. branding) more than other media. It would be helpful to collect large-scale datasets which induce other emotions. However, we show that crowdsourcing responses to online media content is an effective way to train affective content analysis systems [33].

REFERENCES

- [1] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

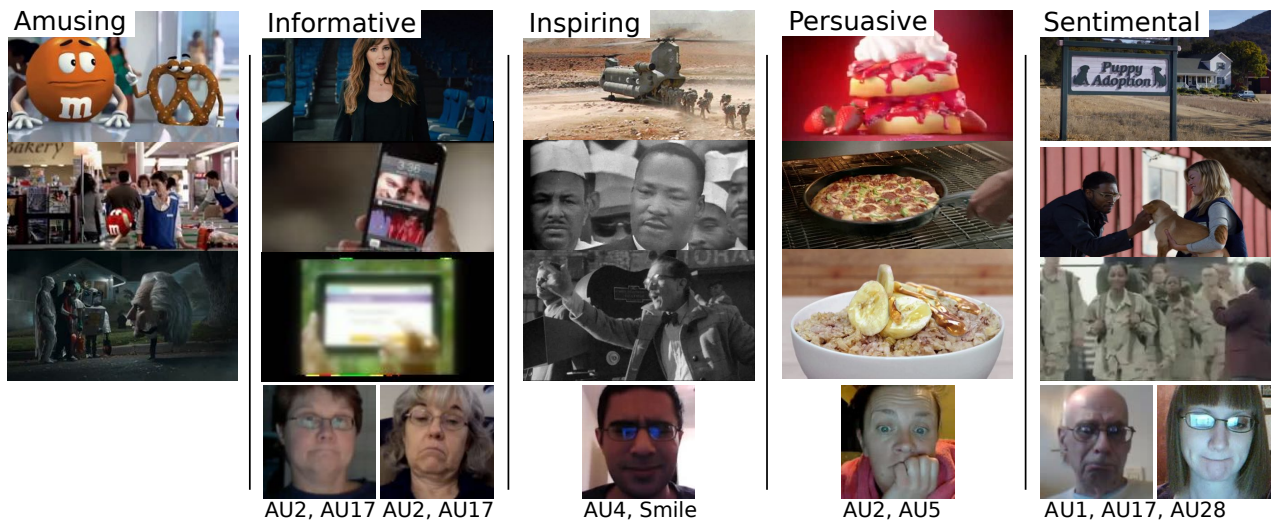


Fig. 5. Examples of frames from the media content in each affect class and facial responses with common AU combinations. Persuasive ads often featured food and were accompanied by expressions with wide eyes and raised eyebrows. Sentimental ads often featured animals or armed forces and were associated with inner brow and chin raising.

- [2] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.
- [3] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 677–680.
- [4] S. Wang and Q. Ji, "Video affective content analysis: a survey of state-of-the-art methods," *Affective Computing, IEEE Transactions on*, vol. 6, no. 4, pp. 410–430, 2015.
- [5] M. Soleymani and M. Pantic, "Human-centered implicit tagging: Overview and perspectives," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3304–3309.
- [6] D. McDuff, R. El Kaliouby, and R. Picard, "Predicting online media effectiveness based on smile responses gathered over the internet," in *Automatic Face and Gesture Recognition, 2013. Proceedings. Tenth IEEE International Conference on*, 2013.
- [7] D. J. McDuff, "Crowdsourcing affective responses for predicting media effectiveness," Ph.D. dissertation, Massachusetts Institute of Technology, 2014.
- [8] D. McDuff and R. el Kaliouby, "Applications of automated facial coding in media measurement," *IEEE Transactions on Affective Computing*, in press.
- [9] H. Joho, J. Staiano, N. Sebe, and J. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools and Applications*, pp. 1–19, 2011.
- [10] D. McDuff, J. M. Girard, and R. el Kaliouby, "Large-scale observational evidence of cross-cultural differences in facial behavior," *Journal of Nonverbal Behavior*, pp. 1–19, 2016.
- [11] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.
- [12] C. E. Izard, "Innate and universal facial expressions: Evidence from developmental and cross-cultural research," *Psychological Bulletin*, vol. 115, no. 2, pp. 288–299, 1994.
- [13] J. A. Russell, "Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies," *Psychological Bulletin*, vol. 115, no. 1, p. 102, 1994.
- [14] H. Gunes and H. Hung, "Is automatic facial expression recognition of emotions coming to a dead end? the rise of the new kids on the block," *Image and Vision Computing*, 2016.
- [15] P. Ekman and W. Friesen, *The Facial Action Coding System (FACS)*. Consulting Psychologists Press, Stanford University, Palo Alto, 1978.
- [16] T. Senechal, D. McDuff, and R. Kaliouby, "Facial action unit detection using active learning and an efficient non-linear kernel approximation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 10–18.
- [17] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao, "Facial affect 'in-the-wild,'" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 36–47.
- [18] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video Emotion Recognition with Transferred Deep Feature Encodings," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16*. New York, New York, USA: ACM Press, 2016, pp. 15–22.
- [19] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus Development for Affective Video Indexing," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1075–1089, jun 2014.
- [20] M. Pantic and A. Vinciarelli, "Implicit Human-Centered Tagging," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 173–180, nov 2009.
- [21] M. Soleymani, S. Asghari Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, jan 2016.
- [22] F. Silveira, B. Eriksson, A. Sheth, and A. Sheppard, "Predicting audience responses to movie content from electro-dermal activity signals," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13*. New York, New York, USA: ACM Press, 2013, p. 707.
- [23] D. McDuff, "Discovering facial expressions for states of amused, persuaded, informed, sentimental and inspired," in *Proceedings of the 18th international conference on Multimodal Interaction*. ACM, 2016.
- [24] D. McDuff, R. El Kaliouby, and R. Picard, "Crowdsourcing facial responses to online videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 456–468, 2012.
- [25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, apr 2016.
- [26] F. Eyben, F. Wening, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia - MM '13*. New York, New York, USA: ACM Press, oct 2013, pp. 835–838.
- [27] S. Bakhshi, D. Shamma, L. Kennedy, Y. Song, P. de Juan, and J. Kaye, "Fast, Cheap, and Good: Why Animated GIFs Engage Us," in *Proceedings of the 34rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, 2016.
- [28] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *International Conference on Image Processing*, vol. 1. IEEE, 2002, pp. 1–477.
- [29] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang, "Visual affect around the world: A large-scale multilingual visual

- sentiment ontology,” in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 159–168.
- [30] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. Proceedings of the IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. 1–511.
- [31] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, “Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit,” in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2016, pp. 3723–3726.
- [32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [33] R. R. Morris and D. McDuff, “Crowdsourcing techniques for affective computing,” *The Oxford Handbook of Affective Computing*, p. 384, 2014.