Multimodal Anticipated versus Actual Perceptual Reactions

Monali Saraf
University of Maryland, Baltimore County
Baltimore, Maryland
monali2@umbc.edu

Tyrell Roberts
Colgate University
Hamilton, New York
troberts@colgate.edu

Raymond Ptucha
Rochester Institute of Technology
Rochester, New York
rwpeec@rit.edu

Christopher Homan
Rochester Institute of Technology
Rochester, New York
cmh@cs.rit.edu

Cecilia Ovesdotter Alm
Rochester Institute of Technology
Rochester, New York
coagla@rit.edu

ABSTRACT

We introduce an experimental method where subjects watch and rate humorous versus neutral videos while their reactions are collected in three modes: non-linguistic verbalizations (laughter), facial expressions, and skin response. We use unimodal analysis and predictive modeling to examine the relationship between the reactions and the perceptions anticipated by experimenters versus the subjects' reported actual ones. We find expected associations for facial expressions and amusement, but not skin response. Laughter, while relatively infrequent, strongly indicates amusement when it

ICMI '19 Adjunct, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6937-4/19/10.
https://doi.org/10.1145/3351529.3360663

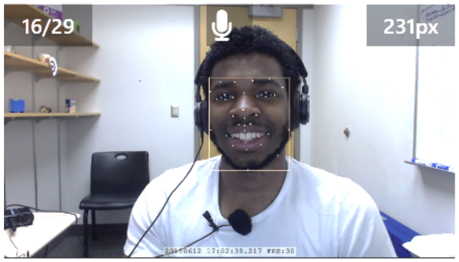
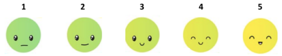


Figure 1: Video and microphone capture of facial and spoken reactions; headphones are used to isolate video sound.

Table 1: Questions that were asked in each post-clip survey.

Which face best describes your reaction to the most recent clip? 
Would you recommend this video to a friend?

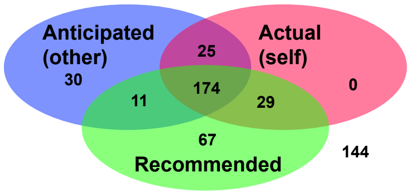


Figure 2: Overlap of annotations of other-annotated, anticipated amusement versus self-annotated actual amusement, as well as subjects' recommendation of videos to a friend.

occurs. Our method can apply generally for comparing anticipated versus actual responses when collecting data for learning affective human response.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Human-centered computing** → *User studies.*

KEYWORDS

multimodal, actual versus anticipated reactions, amusement

ACM Reference Format:
Monali Saraf, Tyrell Roberts, Raymond Ptucha, Christopher Homan, and Cecilia Ovesdotter Alm. 2019. Multimodal Anticipated versus Actual Perceptual Reactions. In *Adjunct of the 2019 International Conference on Multimodal Interaction (ICMI '19 Adjunct), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3351529.3360663

INTRODUCTION AND RELATED WORK

For the supervised learning of affective reactions, what is the impact of using labels from those responding in the moment with visceral, affective reactions (i.e., *self-annotation*) versus from those, such as experimenters, who seek an authoritative label for each item (i.e., *other-annotation*)? We focus on analysis and machine learning-based automated prediction that uses labels for affective experiences, representing either the experimenters' other-annotated (anticipated) or the subjects' self-annotated (actual) cognitive perception of amusement. We then predict both using the subjects' multimodal biophysical data–facial expressions, spoken reactions, galvanic skin response (GSR)–from when they viewed video clips (Figure 1).

We study amusement because of its positive impact on humans' well-being. Additionally, the risk to subjects when inducing this affective state is low and may even have benefits [4, 12]. The choice was also motivated by the feasibility of eliciting amusement in a laboratory setting [5].

Prior unimodal studies that study positive emotions [2, 6–8, 10, 11] motivate our multimodal approach. Our approach addresses fine-grained analysis of anticipated versus actual amusement.

PROCEDURE

All 30 subjects (16 males, 14 females, ages 19-34) were students from a large US university and reimbursed $14. A plurality were computing and information sciences majors (46%); the next largest group (18%) studied humanities or arts. More than 40% consumed more than 10 hours of media content a week and 29% chose comedy as their preferred television and movie genre.

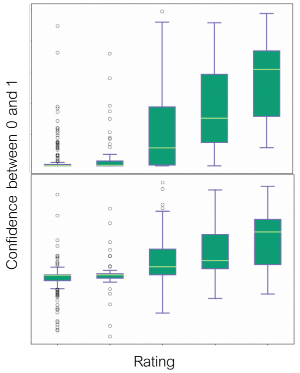


Figure 3: Five-point self-rankings for clips plotted by mean confidence (0-1) of face estimates for smile (top) or valence (bottom).

Table 2: Top-5 features that, if used alone, had top accuracy in leave-one-subject-out cross validation (similar for leave-one-video-out). J = joy, S = smile, C = cheek raise, E = engagement, V = valence, mx = max, mn = mean, sd = stand dev.

P1	P2	P3
J mx 0.81	**J mx** 0.84	J mn 0.73
S mx 0.80	S mx 0.83	E mx 0.71
C mx 0.80	J mn 0.83	J sd 0.71
C mn 0.79	**S sd** 0.82	**S sd** 0.71
S sd 0.79	E m 0.81	**C mn** 0.70

Table 3: Accuracy per modality. As expected, the face outperformed other modalities. Laughter improved over GSR on the self-annotated P2, and on P3, but not on other-annotated P1. This indicates difficulty in anticipating others' strong amusement.

Modality	P1	P2	P3
Face	0.77	0.82	0.68
GSR	0.32	0.49	0.37
Laughter	0.11	0.52	0.56

In this IRB-approved study, subjects watched a randomized-per-subject series of video clips selected by the researchers. Each was followed by a *post-clip survey* (Table 1), and an *emotional palate cleanser task*. The task was designed to be completed in 5–15 seconds and bring the emotional state of subjects back to a baseline position. It involved answering simple multiple choice questions for images.

There was an equal number of other-annotated clips anticipated to be amusing versus neutral clips. The latter were intended not to evoke an emotional response. To prevent disengagement, clips lasted 20–60 seconds.

We used iMotions [9] to extract the *confidence* (here meaning the likelihood of a given facial expression being experienced) of estimated facial expressions and face-estimated emotions in real time. A Logitech C922x Pro Stream Webcam recorded the face at a resolution of 1080p. For capturing laughter reactions, we used the Audacity Cross-Platform Sound Editor [1]. We calibrated the sound levels for each subject. For GSR, we used a Shimmer3 GSR+ device on the middle phalanx of the index and middle fingers of the non-dominant hand. Due to sensor sensitivity, we asked subjects to keep their hand still.

We removed ambient noise, compressed the volume, and adjusted and balanced the frequency responses of each audio recording. We split each subject into subintervals per task: video, post-clip survey, and emotional palate cleanser task. We performed semi-automatic audio analysis and post-corrected inaccurate intervals of laughter after they were automatically detected in Praat [3], and processed acoustic features such as laughter intensity and proportion. Using iMotions, we estimated the confidence of emotional facial expressions (Joy, Anger, Fear, Contempt, Disgust, Sadness, and Surprise) and facial behaviors such as cheek raise or engagement. We applied a median and a low-pass filter on the values extracted for facial behaviors and GSR. The set for analysis and machine learning comprised 480 instances of multimodal reactions (16 video clips per subject).

RESULTS AND DISCUSSION

We consider three binary prediction problems:
[P1] Predicting anticipated, other-annotated amusement versus neutral labels.
[P2] Predicting actual, self-annotated amusement after binarizing replies (Table 1) into 1-2 versus 3-5 to compensate for subject-to-subject scale interpretation (given the ratings' empirical distribution).
[P3] We predicted whether or not a subject would recommend a video to a friend, or not.

Figure 2 shows that self-annotation may not result in the same labels as other-annotation. For example, of the 240 other-labeled clips anticipated to induce amusement, 83% of subjects also self-labeled humorous stimuli as amusing. If also considering recommending a clip to a friend, the number decreased to 73%. Further, no subjects self-annotated stimuli as amusing unless the clip was either anticipated to be so, or if they would recommend it to a friend.

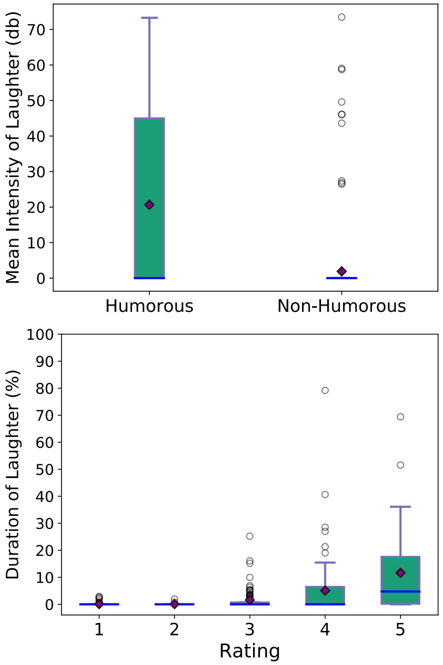


Figure 4: Top: Mean laughter intensity differs for anticipated amusement versus neutral stimuli. Outliers show amusement may occur with neutral stimuli. Bottom: Proportion laughter increases with higher amusement ratings.

Table 4: Top-3 features with lowest p-values per modality. They are lowest for face features.

Mode	Features	t	p
Face	joy sd	19.31	<0.001
	smile sd	18.99	<0.001
	cheek raise max	18.75	<0.001
GSR	SC max	1.69	0.09
	SC min	-1.44	0.15
	SC sd	1.43	0.15
Laughter	dur mn	-1.40	0.16
	duration	-1.13	0.25
	# laughs	1.00	0.31

Figure 4 shows that stimuli anticipated to be amusing had higher mean laughter intensity than neutral stimuli and that as proportion of laughter increased, so did subject ratings. Further, subjects tend to recommend clips with higher mean laughter intensity. However, most stimuli intended to be amusing (56%) had no traces of laughter, and so statistical tests (Table 4) were less relevant for laughter.

Regarding facial and skin response features, to determine the likelihood that response features for videos anticipated to be amusing had the same underlying distributions as neutral ones, we applied two-sample t-tests. These indicated the strongest differences for iMotions-estimated Joy, Cheek Raising, Smiling, Valence, and Lip Stretching, and weaker ones for Interocular Distance, Attention, and Inner Brow Raise. Table 4 shows the three features with the lowest p-values per modality. Overall, facial features are the lowest. Table 4 shows t-tests for GSR did not differ across the two types of videos. Figure 3 shows that videos rated to be more amusing elicited stronger facial reactions.

We used random forest classifiers for the binary problems: P1 (other-annotated, anticipated amusement), P2 (self-annotated, actual amusement), and P3 (recommendation to a friend), with leave-one-*subject*-out and leave-one-*video*-out cross-validation. Table 2 shows the top features for the former. Predicting self-annotated amusement (P2) performed best in both forms of cross-validation, and recommendation to a friend (P3) lowest. Across problems, in both types of cross-validation, features associated with the face yielded the highest accuracy, as seen in Table 2. When all facial features were used in each prediction problem, the classifier performed better than with GSR or laughter combined, likely as most clips did not elicit laughter.

An ablation study with random forests determined the top-performing feature combinations for P1-3 using, first, leave-one-subject-out cross-validation to find the top-performing feature, and then, of the remaining features, discovering the top-performing pair of features. This process was repeated until every feature was included. For P1 the highest accuracy (0.82) was achieved at 23 features, while for P2 the highest accuracy (0.85) used 26 features, and for P3 the highest accuracy (0.76) had 18 features. Results in Table 3 affirmed the key status of facial expressions, followed by laughter for the self-annotated P2 as well as for the recommendation to a friend in P3. Actually laughing likely influenced subjects’ decision to rate stimuli as highly amusing or as a clip they would recommend to a friend. For other-annotated P1, GSR instead followed face, which indicates it is difficult to anticipate other individuals’ strong amusement reactions.

Future work can explore how the various features can be used to predict other emotions.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. IIS-1559889. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Audacity Team. 2017. Audacity(R): Free Audio Editor and Recorder [Computer application]. *Version 2.2* (2017).

[2] Oswald Barral, Ilkka Kosunen, and Giulio Jacucci. 2018. No Need to Laugh Out Loud: Predicting Humor Appraisal of Comic Strips Based on Physiological Signals in a Realistic Environment. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 6 (2018), 40.

[3] Paul Boersma and David Weenink. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.42, retrieved summer 2018. http://www.praat.org/

[4] Amy Drahota, Alan Costall, and Vasudevi Reddy. 2008. The vocal communication of different kinds of smile. *Speech Communication* 50, 4 (2008), 278–287.

[5] Nicole R Giuliani, Kateri McRae, and James J Gross. 2008. The up-and down-regulation of amusement: experiential, behavioral, and autonomic consequences. *Emotion* 8, 5 (2008), 714.

[6] Raj Kumar Gupta, Prasanta Bhattacharya, and Yinping Yang. 2019. What Constitutes Happiness? Predicting and Characterizing the Ingredients of Happiness Using Emotion Intensity Analysis. *CEUR Workshop Proceedings* 2328 (2019).

[7] M. Hoque and R. W. Picard. 2011. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *Face and Gesture 2011*. 354–359. https://doi.org/10.1109/FG.2011.5771425

[8] M. E. Hoque, D. J. McDuff, and R. W. Picard. 2012. Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles. *IEEE Transactions on Affective Computing* 3, 3 (July 2012), 323–334. https://doi.org/10.1109/T-AFFC.2012.11

[9] iMotions Team. 2018. iMotions Biometric Research Platform. *Version 7.1* (2018).

[10] Daniel J. McDuff. 2016. Discovering facial expressions for states of amused, persuaded, informed, sentimental and inspired. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*. 71–75. https://doi.org/10.1145/2993148.2993192

[11] Jordan Edward Shea, Cecilia O. Alm, and Reynold. Bailey. 2018. Contemporary multimodal data collection methodology for reliable inference of authentic surprise. In *Proceedings of the 2018 IEEE Western New York Image and Signal Processing Workshop*. IEEE.

[12] Vivien C Tartter. 1980. Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & psychophysics* 27, 1 (1980), 24–27.