Considerations for Face-based Data Estimates: Affect Reactions to Videos

Gustaf Bohlin[1], Kristoffer Linderman[1], Cecilia Ovesdotter Alm[2] and Reynold Bailey[2]
[1]*Malmö University, Sweden*
[2]*Rochester Institute of Technology, U.S.A.*

Keywords: Affective Reactions, Facial Expressions Estimates, Face-based Pulse Estimates.

Abstract: Video streaming is becoming the new standard for watching videos, providing an opportunity for affective video recommendation that leverages noninvasive sensing data from viewers to suggest content. Face-based data has the distinct advantage that it can be collected noninvasively with minimal equipment such as a simple webcam. Face recordings can be used for estimating individuals' emotional states based on their facial movements and also for estimating pulse as a signal for emotional reactions. We provide a focused case-based contribution by reporting on methodological challenges experienced in a research study with face-based data estimates which are then used in predicting affective reactions. We build on lessons learned to formulate a set of recommendations that can be useful for continued work towards affective video recommendation.

1 INTRODUCTION

Face-based data has the distinct advantage that it can be collected noninvasively with minimal equipment such as a simple webcam. Face recordings can be used to estimate an individual's emotion based on their facial movements. Pulse can also be estimated from videos of the face, providing additional cues about the viewers' emotional state.

We provide a focused case study contribution by reporting on methodological challenges with face-based data estimates, experienced in the context of predictive modeling as a step towards affective video recommendation. We captured webcam recordings of users' faces and upper bodies as they watched video clips intended to evoke reactions of anger, fear, happiness, sadness, or surprise. We report on the use of face-based estimates of emotional facial movements and face-based pulse for computational modeling to predict a user's rating of a video, comparing against the explicit self-reported rating.

Video streaming is becoming the new standard for watching videos, providing a need to suggest content to users. Although an individual's rating is valuable to a recommendation system, most people do not rate the videos they watch (for example, 50% of the subjects involved in this study indicated that they never rate videos). This provides an opportunity for affective video recommendation that leverages noninvasive sensing data from viewers to suggest new content. Affective video recommendation sets out to analyze the users' emotions while they are watching videos in order to conclude what to recommend.

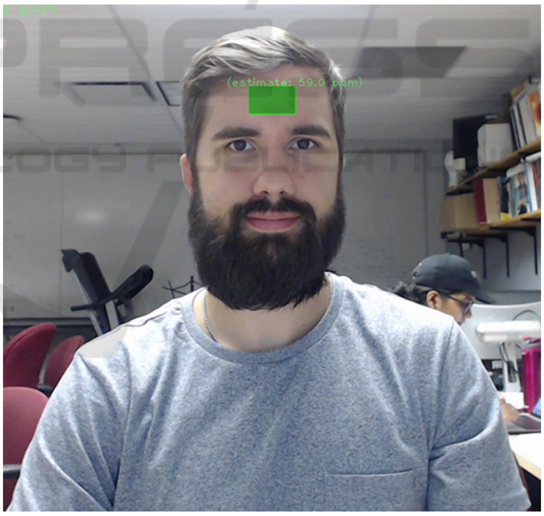

Figure 1: Facial expression and pulse were both captured using standard webcams as demonstrated here by one of the authors. The green box indicates where the pulse was tracked.

Bohlin, G., Linderman, K., Alm, C. and Bailey, R.
Considerations for Face-based Data Estimates: Affect Reactions to Videos.
DOI: 10.5220/0007687301880194
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 188-194
ISBN: 978-989-758-354-4

2 RELATED WORK

Face-based estimates are commonly used in inference of emotional experiences. For example, (Busso et al., 2004) explored multimodal emotion recognition using speech and facial expressions. An actor read sentences while being recorded and face markers were utilized to interpret facial muscular movement. Similarly, (Ioannou et al., 2005) extracted face features and explored the understanding of users' emotional states with a neurofuzzy method and facial animation parameters. As another example, (Tarnowski et al., 2017) used a Microsoft Kinect to record a 3D model of subjects' faces with numerous facial points. They recognized seven emotions using facial expressions, a k-NN classifier, and a neural network . Work in affective computing has also focused on reactions for specific emotions. For instance, (Shea et al., 2018) studied intuitively extracted reactions to surprise, spanning multiple modalities. Estimated facial expressions were particularly important for identifying naturally occurring surprise reactions.

More specifically, affective computing methods have been considered promising for video recommendation and classification. (Zhao et al., 2013) presented a framework for recognizing human facial expressions to create a classifier, which identified what genre viewers watched from their facial expressions. However, as they drew on acted facial data, reactions tended to involve exaggerations rather than corresponding to less direct, more intuitive and natural expressions, resulting in modeling unsuitable for actual practical use.

The use of facial data towards video recommendation was explored by (Rajenderan, 2014). To facial expressions, Rajenderan added analysis of the pulse modality, calculated with a method called photoplethysmography, developed at MIT by (Poh et al., 2011). The work was continued by (Diaz et al., 2018), with a focus on estimating and visualizing viewers' dominant emotions over the course of a video. The photoplethysmography method uses fluctuations in skin color related to blood volume and the proportion of reflected light to help estimate the viewer's pulse. For this case study, we also apply photoplethysmography for non-invasive pulse estimation (see Figure 1), with recalibration occurring between each video viewed by the subject in the study.

3 METHODS

Conducting an experiment that entirely focuses on face-based capture provides an opportunity to reflect on challenges that occur when working with face-based data estimates. We build on this experience in presenting examples that illustrate methodological considerations and summarize lessons learned, as a springboard to formulate a set of recommendations that can be useful for continued work towards affective video recommendation. This section describes how we collected face data and processed the face-based estimates for use in predictive modeling, taking a step towards video recommendation.

3.1 Data Collection

Equipment. The equipment used for data collection included two standard webcams operating in real-time: the Logitech C922 Pro Stream Webcam and the Logitech Pro 9000 Webcam. One webcam was used to capture the subject's facial expressions, while the other was used to estimate the subject's pulse using the aforementioned photoplethysmography method. Additional hardware included a desktop computer with a 24" computer monitor with external loudspeakers, keyboard, and a mouse.

Stimuli. The experiment included carefully selected short video clips with content from movies, TV programs, or videos. The clips intended to elicit reactions corresponding to five major emotions: happiness, sadness, anger, fear, and surprise. The emotional impact of the videos was assessed jointly by the authors. Three clips were included per emotion category with a total of 15 video clips. We avoided content that might cause strong discomfort. Table 1 provides an example from each emotion category. Subjects consented to participating in the IRB-approved study.

Procedure. The data collection process is illustrated in Figure 2. Each subject was given oral instructions explaining the outline of what the experiment would look like. After completing the consent form and receiving a walk-through of the experiment, participants filled out a demographic survey. They did not know what clips they were watching prior to viewing the videos.

(Diaz et al., 2018) discussed that an experimenter being present could potentially have an effect on a viewer's emotional expressions. Accordingly, the subject was alone in the room during the experiment to mitigate any such an effect. For each video:

1. the subject was shown an image with instructions for the pulse calibration
2. a 50 second video of a countdown was shown in order to calibrate the pulse estimation

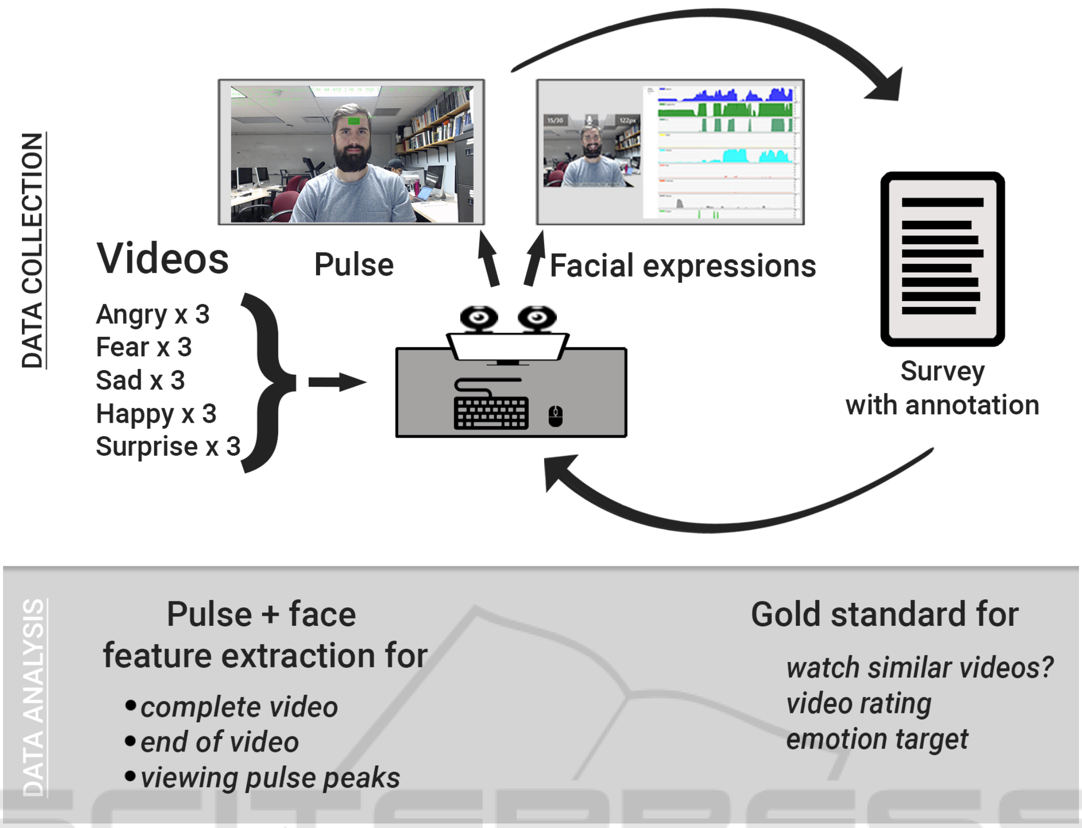

Figure 2: Data collection procedure. Subjects viewed fifteen videos to elicit affective reactions from five emotion categories.

3. the video was shown
4. the subject filled out a survey regarding the video they watched

This was repeated 15 times, and each subject watched all videos. The order of the videos displayed was randomized for each subject. The survey after each video included the following questions, adapted from (Diaz et al., 2018):

- Have you seen this video before?
- What did you feel when watching the video?
- On a scale of 1 to 5, how would you rate the video?
- Would you want to watch similar videos? (Yes, No, Maybe)

For the second question, the subject could choose any one or more of the following emotions as applicable: *happiness, sadness, anger, fear, surprise*, and *other (please specify)*. At the end of the experiment the subjects were thanked for their participation and received a cash payment of $12 USD.

The facial expressions were processed using Affectiva ((McDuff et al., 2016)) in iMotions, focusing on automatically inferred high-level facial expressions such as joy and anger, as shown in Table 2. All features extracted from iMotions were represented as a numeric value reflecting the confidence that the feature was expressed. Every feature extracted was then aggregated in five ways (min, max, average, median, and standard deviation) for subsequent modeling analysis.

The estimated pulse was processed into three types of features: (1) pulse derivative, or the change in pulse from one sample to the next; (2) absolute pulse derivative, meaning the absolute value of the pulse derivative, and (3) pulse derivative direction represented as 1 (increasing), 0 (no change), or -1 (decreasing). We used measures of change as opposed to the exact estimated pulse values because of differences in pulse between individuals as well as concerns about inaccurate values; by focusing on measures of change we mitigated such issues and centered on trends instead.

Subjects. The data collection involved 32 volunteers (17 female and 15 male) recruited on campus through study announcements. Twenty-six reported an age between 18-24 and six reported an age between 25-44. In the quantitative analysis, five subjects were excluded because of data quality concerns.

Table 1: Examples of the video clips used to elicit emotional reactions. Length of clip in parentheses.

Happy	Sad	Anger	Fear	Surprise
Despicable me 2	**Marley & me**	**Witness**	**Shining**	**Magic show**
A man has a date with a woman that goes well. The day after, he is happy and dances around the town. (1:57)	A dog is being put down. Flashbacks of the dog's life in a happy family are shown. (2:02)	A group of Amish people are entering a town. When they enter a gang of youths harass them as they are unwilling to fight back. (1:16)	A video of a frightened boy followed by a slow pan through an empty living room paired with menacing sound. (1:22)	A man is performing magic where he produces birds out of nowhere. In the end he also reveals a woman that could not be seen during the performance. (2:01)

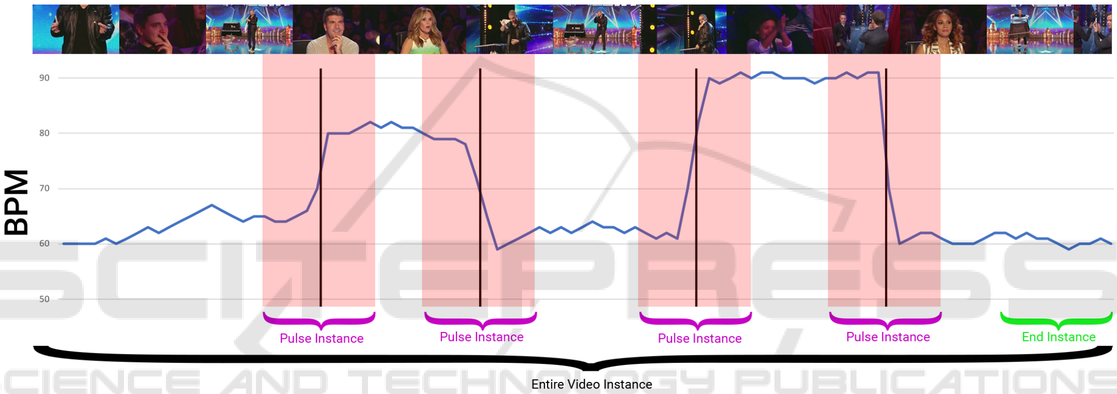

Figure 3: Theoretical diagram explicating three approaches to consider data from the viewing process.

Table 2: Facial expression features.

Anger	Sadness	Disgust
Joy	Surprise	Fear
Contempt	Smile	

Computational Modeling. As a step towards affective video recommendation, and given the modest size of the dataset, we used a Support Vector Classifier (SVC) from scikit-learn[1] to implement predictive modeling of the subjects' ratings of the videos (5 classes) and whether they would watch similar videos again (3 classes). For the machine learning model two types of face-based estimated feature modalities were used: facial expressions and pulse. We used ablation to tune the model to well-performing features. For every model trained, the accuracy was calculated using an average of the score for all folds where each fold left one subject out. We also explored Decision Tree and Random Forest methods, and we compared against a baseline classifier from scikit-learn.

To investigate which data from the viewing process enabled prediction, we considered three approaches in the feature aggregation, as shown in Figure 3, considering: (1) the entire video, (2) 10 second windows anchored in time points where viewers demonstrated a significant change in their estimated pulse (with 5 seconds before and after), and (3) the 10 last seconds of the clip where the clip highlight tended to occur.

4 FINDINGS

We first include example findings from analyzing the data from the participants that shed light on methodological challenges with face-based capture. Second, we provide results from the classification developed based on the face-based estimated features. We also discuss limitations of this study.

[1] https://scikit-learn.org

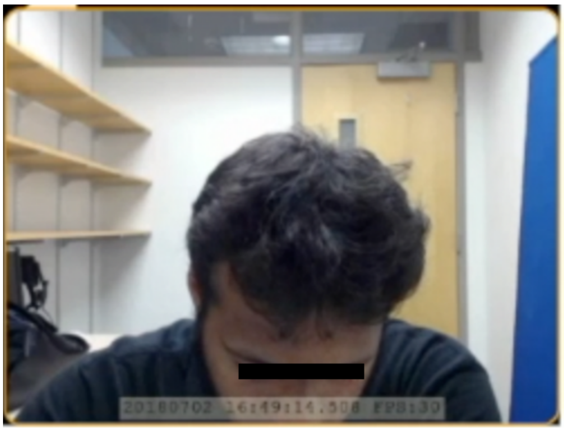
[redacted]

Figure 4: A viewer looks away.

4.1 Examples of Challenges

The following examples are indicative of challenges encountered when affective estimates are based merely on face-based capture.

Example 1: Looking Away. Subjects were not monitored by a person in the room. Several subjects did at some point in the data collection process begin to look around the room or at their phone. When this happened the pulse estimation lost track of their face and it also temporarily obstructed the facial expression processing; see Figure 4. A contributing reason could be leaving subjects alone in the experimentation room.

Example 2: Abundance of Joy. While more prominent for happy videos, facial indicators of joy occur for videos of various emotion categories, as shown in Figure 5. We suspect that this is due to subjects smiling when experiencing other emotions than just joy. For instance, they could be smiling as a sign of frustration or smiling at something nice in a sad scene. (Hoque et al., 2012) reported on a study which explored how smiles occurred with such noncanonical emotional reactions.

Example 3: Other Visual Reactions. There are strong visual cues for emotional reactions that extend beyond facial movements, which a human could recognize immediately, but that may be missed by facial expression analysis focusing on facial movements. Figure 6 shows a subject shedding tears while viewing a clip intended to elicit sadness; the analysis based on facial movements detected only a small amount of sadness for short periods of time as indicated by the circled episodes in Figure 6.

4.2 Subpar Predictive Modeling with Face-based Estimations

The results of the ablation and in turn the best performing classifiers are in Table 3. Face-based features from across the entire clip appear to generate more accurate models, however, the subpar prediction performance (only slight improvement over the comparison baseline) suggests that sole reliance on face-based features, which suffer from methodological challenges during capture or processing, did not aid robust prediction, at least not in this case.

There are also other issues with face-based estimates. For instance, one face experienced repeated track loss even though the subject remained still, highlighting nonrobustness to the range of faces.

Limitations. The videos used were intended to elicit emotional reactions yet were intentionally mild to mitigate emotional triggers, and they were at most 3 minutes long, which may have resulted in less emotional expression or absence of such reactions. In addition, 27 participants represents a modest sample size with implications for the effectiveness of the machine learning modeling.

5 DISCUSSION

This case study identified challenges for face-based data estimates with implications for producing reliable data, data analysis, and predictive modeling of affective reactions, summarized here:

1. Users may not face the camera or their faces may be obstructed and not capturable.
2. Users often multitask and distribute their attention which limits face-based estimation.
3. Models are biased towards expected behaviors and fail to identify reactions when users behave unexpectedly.
4. Models do not yet robustly account for the full range of human diversity.

We recognize that several scenarios need to be explored further such as what happens when multiple faces are tracked in a group simultaneously as well as the impact of lightning conditions.

6 CONCLUSION

Affective video recommendation is an emerging field. While face-based data is an intuitive, unobtrusive mo-

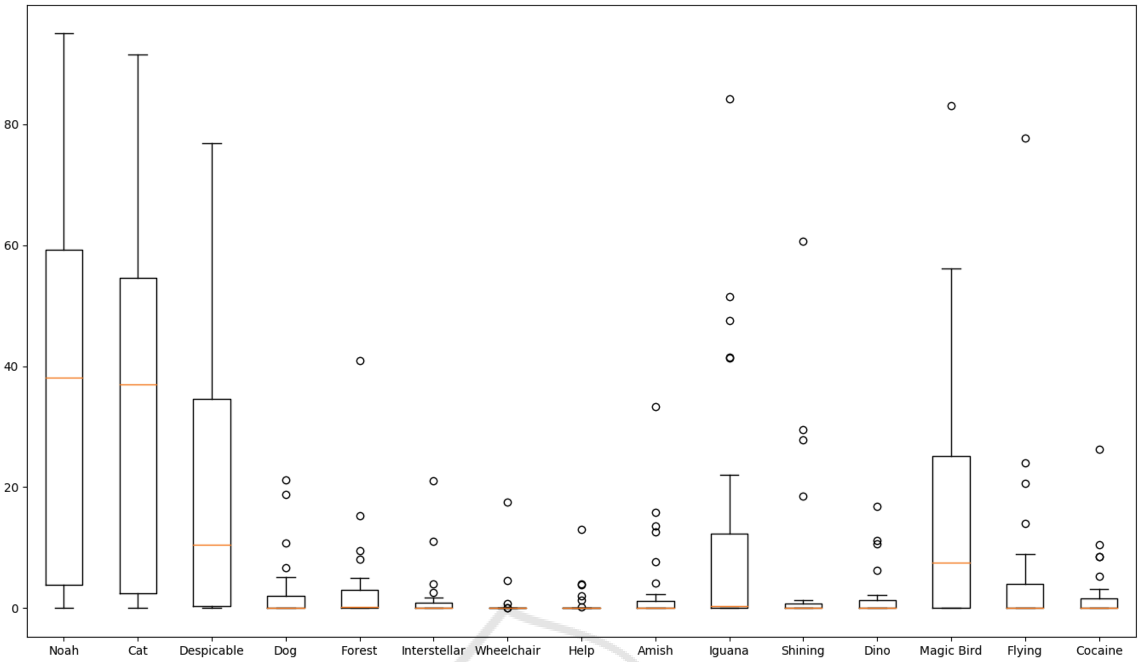

Figure 5: Abundance of joy. While happy videos are clearly marked by the estimated smiles, so are other categories too.

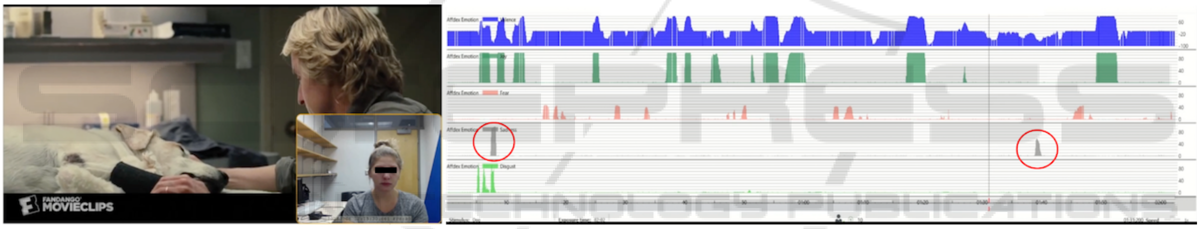

Figure 6: Tears shed when watching a sad clip as visual cue going beyond analyzed facial movements.

Table 3: Mean accuracy from leave-one-subject-out evaluation using SVC with ablation.

			Mean Accuracy		
Classifier	Question	Entire video	End instances	Pulse instances	No. labels
SVC	Which rating?	**36%**	33%	31%	5
Baseline	Which rating?	27%	28%	30%	5
SVC	Watch similar videos?	**48%**	47%	47%	3
Baseline	Watch similar videos?	46%	46%	47%	5

dality to consider for this application, methodological challenges introduce complications, as illustrated in this case study. To set the path to begin to respond to these challenges we formulate three recommendations towards human-aware affective video recommendation. First, system should detect loss of attention measurement and adapt when needed to raise attention with visual or audio cues. Second, robust systems must also leverage multimodal sources of human behavioral data when face-based estimates fail to provide adequate input. Third, face-based software models must be trained with large and diverse sample sizes, accounting for unexpected and uncooperative behaviors.

ACKNOWLEDGEMENTS

We are grateful to Malmö University and Rochester Institute of Technology for their support of the student authors. 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

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pages 205–211, New York, NY, USA. ACM.

Diaz, Y., Alm, C., Nwogu, I., and Bailey, R. (2018). Towards an affective video recommendation system. In *Workshop on Human-Centered Computational Sensing at PerCom*, pages 137–142.

Hoque, M. E., McDuff, D. J., and Picard, R. W. (2012). Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing*, 3(3):323–334.

Ioannou, S., Raouzaiou, A., A Tzouvaras, V., Mailis, T., Karpouzis, K., and Kollias, S. (2005). Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural networks : the official journal of the International Neural Network Society*, 18:423–35.

McDuff, D., Mahmoud, A. N., Mavadati, M., Amr, M., Turcot, J., and Kaliouby, R. E. (2016). Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. In Kaye, J., Druin, A., Lampe, C., Morris, D., and Hourcade, J. P., editors, *CHI Extended Abstracts*, pages 3723–3726. ACM.

Poh, M.-Z., McDuff, D. J., and Picard, R. W. (2011). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11.

Rajenderan, A. (2014). An affective movie recommendation system. Master's thesis, Rochester Institute of Technology.

Shea, J. E., Alm, C. O., and Bailey., R. (2018). Contemporary multimodal data collection methodology for reliable inference of authentic surprise.

Tarnowski, P., Koodziej, M., Majkowski, A., and Rak, R. J. (2017). Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175 – 1184. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

Zhao, S., Yao, H., and Sun, X. (2013). Video classification and recommendation based on affective analysis of viewers. *Neurocomputing*, 119:101–110.