From Dials to Facial Coding: Automated Detection of Spontaneous Facial Expressions for Media Research

Evan Kodra, Thibaud Senechal, Daniel McDuff, Rana el Kaliouby

Abstract— **Typical consumer media research requires the recruitment and coordination of hundreds of panelists and the use of relatively expensive equipment. In this work, we compare results from a legacy hardware dial mechanism for measuring media preference to those from automated facial analysis on two television programs, a sitcom and a drama series. We present an automated system for facial action detection as well as a continuous measure of valence. The results demonstrate that automated facial analysis provides similar as well as additional insights on moment-to-moment affective response in a way that is unobtrusive, scalable and practical. Specifically, highly significant correlations are found between the dial and facial expression data. For specific moments where the two methods disagree, facial expression data provides additional traceable insights that cannot be obtained from dial data. Furthermore, this data can be obtained at a fraction of the cost; in this work, the facial expression data panel size is only about 5% of the sample size needed to obtain reliable dial data. Results have substantial implications for the future of media research and audience measurement.**

I. Introduction

Affect valence is the positive or negative emotional charge of an event or experience [18]. The need to quantify valence arises in numerous domains including in consumer media research. Television shows and movies aim to interest and entertain millions. Before a TV show or movie is launched, content providers often test this content using a consumer panel. The results are used to optimize the content (e.g., tweak the choice and duration of scenes) or assess engagement with certain characters in the show. With one popular measurement method [17], panelists are asked to turn a hardware dial to quantify valence throughout the show. In this study, dial values (collected at discrete time intervals of 1 Hz) range from 0 to 100, where panelists are told that 0 is disinterest, 50 neutral, and 100 interest in the show.

While the dial approach has been used quite successfully for many years [17], it has several drawbacks. First, its use requires the recruitment and coordination of panelists in person, typically in a limited geographic and demographic locale. Second, large consumer panels are needed to produce reliable aggregate continuous measurement of affective responses. Third, analysis is limited to one dimensional dial data [17]: e.g., testers would know that valence is negative, but would not have a sense of the viewers specific emotional state. Finally, and perhaps most importantly, are the effects caused by asking a panelist to turn the dial while engaging with content. Some may experience heavy cognitive load from having to manipulate a dial, which may distract them from the media experience and add noise to their response that is difficult to quantify. Specifically, the act of labeling affect impacts the affective response of an individual [9]. Others may become engrossed in the media experience and forget to turn the dial when their affective state changes. In both cases the reports will not match their true state.

The human face is a powerful channel for communicating valence as well as a wide gamut of emotion states. The Facial Action Coding System (FACS) [4], [2] is a comprehensive catalogue of unique action units (AU) that correspond to each independent motion of the face. FACS enables the measurement and scoring of facial activity in an objective, reliable and quantitative way, and is often used to discriminate between subtle differences in facial motion [8]. Facial behavior has been used to measure the effectiveness of media content, typically in the form of short advertising video clips [7], [21]. The general expressiveness of viewers correlates strongly with memory for the content [7], a key measure of success for media content.

One question that is frequently asked, is whether self-report matches nonverbal expressions of emotion such as those on the face. In this work we present an automated method for detecting facial actions, as well as continuous measures of higher-level affective/expression states in spontaneous facial videos. We explore the congruence between the facial measures with continuous measures of self-reported dial data. We demonstrate that automatically coded facial responses of panelists correlate highly with self-reported dial data, but provide additional, deeper insights on moment-to-moment affective responses compared to dial measures. We show that some of the drawbacks of traditional self-report methods such as dials may be overcome or minimized with the use of passive facial expression analysis. Our system can be deployed over the Internet to anyone with a computer, webcam and Internet access. This makes it a highly scalable alternative. Previous work has demonstrated the ability to collect facial responses to online videos from across the world in short amounts of time and with no recruitment necessary [13], [14].

The main contributions of this work are:

1) A system for the automated detection of facial action units including asymmetric ones, as well as a continuous measure of valence in spontaneous facial videos;

E. Kodra, T. Senechal, R. el Kaliouby are with Affectiva Inc., Waltham, MA, USA. {evan.kodra, thibaud.senechal, kaliouby}@affectiva.com

D. McDuff and R. el Kaliouby are with the MIT Media Lab, Cambridge, MA. {djmcduff,kaliouby}@media.mit.edu

2) The first systematic evaluation of self-reported dial data compared to automatically detected facial responses;
3) An exploration of the sample size needed for such studies;
4) A method for finding the moments where self-report and facial expression data diverge, providing additional insights over self-report alone.

In the remainder of this paper we describe: the automated detection of facial actions and valence from video sequences; the collection of dial data and content tested and the insights gained from automated facial analysis and how they complement and improve upon dial measures.

II. Related Work

McDuff et al. [12] show that it is possible to predict affect dial measures of valence based on hand labeled FACS data. However, manual FACS labeling is time consuming, expensive and hard to scale across time and population sizes (e.g., for 30 minute TV programmes manual FACS coding would be very laborious and infeasible for many viewers).

Machine facial coding systems have been developed to help automate coding of responses. The majority of automated facial analysis systems detect facial action units or discrete emotional states (typically six states: amusement, fear, anger, disgust, surprise and sadness). De La Torre and Cohn [23] present a summary of state of the art approaches to automated AU detection. A number of approaches [10], [14] use the assumption that the probability estimate from the classifiers (e.g. distance from the SVM hyper-plane) correlate with the intensity of the action unit or expression.

Continuous prediction of affective states has received a lot of attention recently, particularly the valence and arousal dimensions of the affect circumplex space [6], [16], [19]. The Semaine corpus is a rich multimodal dataset, including video, of emotionally colored interactions between a participant and an operator [15]. The emotion labels for the Semaine dataset were annotated by observers and were not self-reported by the participant. Thus, it is difficult to ascertain whether these labels are correlated with the participant's internal state.

In this paper we present a system for the automated detection of facial expressions (AU02, AU04, AU09/AU10, smile/AU12, asymmetric AU12/AU14, all defined in Table I), as well as a continuous valence classifier. We show how the dynamics of these measures predict the interest of viewers to two full length TV programs and improve upon self-report measures currently used. In addition, the results support the assumption that probability estimates from the classifiers correlate with the intensity of the reported state.

III. Action Unit and Emotion Classification

Our automated facial analysis system detects five facial actions or action combinations (AU02, AU04, AU09/AU10 and smile/AU12, asymmetric AU12/AU14), which are summarized in Table 1. AU09/AU10 is the presence of either of these action units and smile is the presence of AU12 without AU04 or AU09. Asymmetric AU12/AU14 is the presence on unilateral AU12 or unilateral AU14. For simplicity, we will refer to these as action units in this paper. However, some are strictly combinations of FACS AUs. The output of each classifier is a probability estimate of the presence of each action. To detect AUs on a face video, the Nevenvision facial feature tracker (licensed from Google, Inc.) is used to automatically detect the face and tracks 22 facial feature points on each frame of the video, as shown in Figure 2. For each AU, a region of interest (ROI) around the appropriate part of the face is located using the landmark points. The image is cropped to the ROI and a histogram of orientated gradients (HOG) [3] features computed for each region. A random forest classifier, similar to the approach presented in [14] is used to classify the HOG features into AU02, AU04, AU09/AU10. The smile classifier uses a Support Vector Machine (SVM) with RBF kernel as this yielded significantly better performance than the random forest classifier. The automated detection of asymmetric AU12/AU14 is described in detail in [20]. The classification of each action unit classification is treated independently.

In addition to AU detection, we compute Valence (V), a

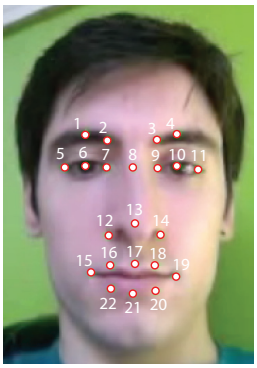

Fig. 2. Locations of the 22 landmark points automatically detected using the Nevenvision tracker.

TABLE I
Definitions and example images for the action units considered in this study. Examples taken from [11].

AU	Description	Example
AU02	Outer corner of eyebrow raised.	
AU04	Eyebrows drawn medially and down.	
AU09	Upper lip raised and inverted; superior part of the nasolabial furrow deepened; nostril dilated by the medial slip of the muscle.	
AU10	Upper lip puller up (either unilaterally or bilaterally).	
AU12/ smile	Lateral lip corner pull without AU04 or AU09.	
AU14	Dimpler (dimples in the cheeks).	

measure for the overall positivity of a persons facial state. V is calculated using HOG features extracted from the whole face, which are then input to a Support Vector Regression (SVR) with RBF kernel. The output ranges from -1 to 1, where -1 indicates negative valence and 1 positive valence.

The classifier outputs are probabilistic and continuous moment-by-moment measures that are computed for each frame of the video. The results reported in this paper suggest that the probabilities are related to the intensity of the underlying state which reinforces other findings [10]. Missing data, which occurs when the facial tracker failed to find a face within a frame, were represented by a row of -1s in the output. From this point onwards these are ignored and treated as null values.

The following section describes how we sourced the spontaneous data used to train and test our system, as well as the labeling and performance of the classifiers.

IV. Training and Accuracy

In order to train the our system, we use a web-based framework similar to that used by McDuff et al. [14] to crowdsource facial videos as people watch online video content. Viewers are given the option to opt-in to turn on their webcam and watch a short video, while their facial expressions are recorded. On the viewers machine, all that is needed is a browser with Flash support and a webcam. The webcam video is streamed to a server at 14 frames a second with a resolution of 320x240. For each video, three FACS trained human labelers coded for the presence or absence of the facial actions: smile (symmetric AU12 with no AU04 or AU09/AU10), AU02, AU04 and AU09/AU10 (either or both of AU09 and AU10 are present), unilateral or asymmetric AU12/AU14. For AU detection, an inter-coder agreement of 50% or higher was required for a positive example to be used for training; 100% agreement on the absence of an AU was required for a negative example.

For valence labeling the following criteria was used:

if (smile present) {valence = +1}
else if (AU04 or AU09 or AU15 present) {valence = -1}
else {valence = 0}

All classifiers were trained and tested with more than 5,000 spontaneous examples labeled by FACS trained human coders. Table II shows the number of positive and negative examples of each AU used in training. The data was divided into three datasets, 50% used for training, 17% used for validation (multiple potential classifiers are evaluated and the best selected) and 33% used for testing. To ensure person-independent experiments, frames from a particular facial video were used exclusively to train or to test the system.

During the validation stage, the HOG parameters, spread of the RBF kernel and size of the facial ROI were varied to find the optimal parameters. The values used in the testing stage were chosen by maximizing the area under the receiver operating characteristic (ROC) curve during validation. Table II shows the area under the ROC curve for each of the action unit classifiers and the valence classifier. For valence, because three classes were used (positive +1, neutral 0 and negative -1), we calculated three ROC curves (positive versus negative labels, positive versus neutral labels and negative versus neutral labels). The mean of the three AUCs is shown in Table II.

TABLE II
Number of videos and frames used for training the action unit and expression classifiers and the area under the ROC curve for testing.

	Videos	**Frames**	**Classifier ROC AUC**
All	2,025	655,000	-
AU02	868	114,000	0.97
AU04	308	16,000	0.72
AU09/AU10	254	58,000	0.84
Smile	93	15,700	0.85
Uni-AU12/14	201	5,100	0.88
Valence	500	65,000	0.90

V. Experiment Description and Method Overview

Two television programmes, a Crime Drama (runtime 40 minutes) and a Sitcom (runtime 20 minutes), provide the two test cases. For the Crime Drama, $n_{cd,d}$ = 184 (83 males, 101 females) panelists were recruited for the dial data, and $n_{cd,f}$ = 11 (6 males, 5 females) were recruited for facial expression analysis. For the Sitcom, $n_{s,d}$ = 216 (88 males, 128 females) panelists were recruited for the dial data, and $n_{s,f}$ = 10 (5 males, 5 females) were recruited for facial expression analysis. With the dial, panelist response is sampled at 1 Hz, while facial videos are processed at 14 Hz, a significantly higher sampling rate sufficient to capture majority of facial behavior dynamics. All outputs from the facial expression analysis are downsampled to 1 Hz to match the dial data. Subsequently, each AU, each expression, and V are smoothed with a polynomial spline to filter out high frequency noise. The spline method was chosen because it is non-causal and hence causes no signal delay at any degree of smoothing. Three degrees of smoothing were attempted; after smoothing each panelists facial expression data, Pearsons linear correlation (ρ) was calculated between the panelist-averaged dial data and each panelist-averaged (smoothed) AU and V. All data analysis was performed in the open source statistical software package R. FIR1 filters are designed and convolved using the function decimate in the add-on package signal.

VI. Results

Table III summarizes correlation results by different order polynomial splines. The smoothing parameter, which can range from 0 to 1, controls how many (proportional to the length of the time series) polynomial terms are used for smoothing. Higher parameters imply more smoothing. To illustrate the motivation for smoothing, we include correlations from the raw metrics (smoothing parameter = none in Table III). Results from a sample of smoothing parameters suggest that correlations are reasonably insensitive to small

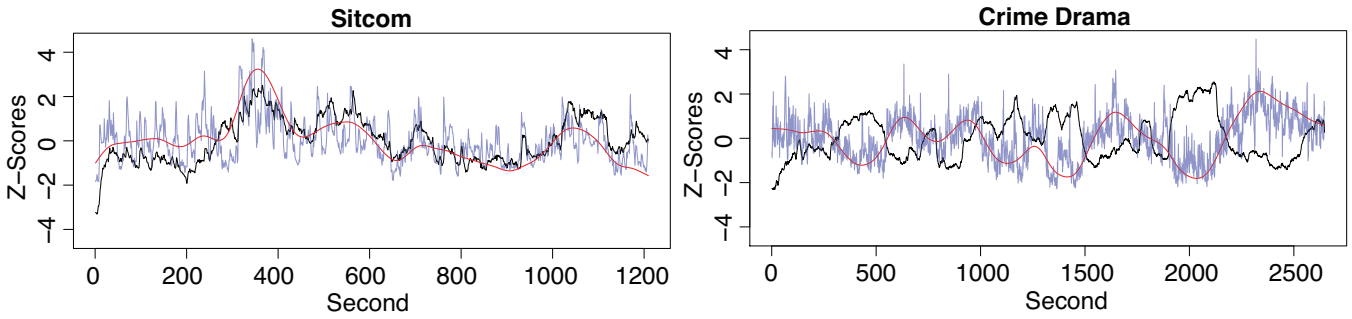

Fig. 3. Time series plot of the dial data (in Black) versus the smoothed metrics (in Red: AU02 for Sitcom and V for Crime Drama). The non-smoothed facial data is shown in faded blue.

TABLE III

PEARSON CORRELATIONS OF THE DIAL AND FACIAL EXPRESSION DATA FOR DIFFERENT VALUES OF SMOOTHING: NONE, 0.3, 0.5 AND 0.7.

Smoothing Parameter	Crime Drama				Sitcom			
AU	**None**	**0.3**	**0.5**	**0.7**	**None**	**0.3**	**0.5**	**0.7**
AU02	-0.587	-0.715	-0.748	-0.742	0.255	0.371	0.428	0.472
AU04	0.392	0.442	0.476	0.452	-0.123	-0.148	-0.149	-0.065
AU09/AU10	0.047	0.06	0.07	0.114	0.514	0.505	0.526	0.545
Smile/12	0.255	0.383	0.456	0.523	0.401	0.612	0.656	0.69
Uni-AU12	-0.043	-0.178	-0.25	-0.295	0	-0.024	-0.031	-0.105
Valence	-0.232	-0.316	-0.368	-0.395	0.443	0.512	0.526	0.516

changes. Here forward, we use the parameter 0.5 in an attempt to smooth noise yet preserve the main signal. In general, smoothing improves the correlation between the dial data and facial expression data, which makes sense given that viewers are less likely to turn the dial to indicate a change in state at a frequency of 1 Hz.

As shown in Table III, for the Crime Drama, the most notable result is the strong inverse correlation (0.7) between the dial time series and AU02 (Figure 3: left), which is strengthened with smoothing. Moderate correlations are found between AU04, AU09, and smile/AU12. For the Sitcom, the most notable result is the strong positive correlation between the dial time series and smile/AU12 (Table III and Figure 3: right). V and AU02 also exhibit moderate correlations with the dial time series. Given that this is a sitcom that is designed to elicit amusement, the high correlations with smile is intuitive.

While the sample size required to capture properties similar to dial data appears to be orders of magnitude smaller, one might be interested in the following question: how small can our sample size be? We test this notion with AU02 and Smile/AU12 for the Crime Drama and the Sitcom, respectively, via a permutation procedure. We average the metrics for each possible combination of panelists, where we let n range from 3 to n. Subsequently, we compute ρ for each such combination and evaluate at what sample size correlation appears to decline. This inflection point may be useful in guiding researchers selection of panel size and may inform the reliability of expression data in general.

Figure 4 displays the results, using boxplots to show the distributions of ρ for different panel sizes. Results imply that with very small sample sizes, the dial data can be approximated. It is notable that despite the fact that panelists may often have different baselines (or mean AU probability biases), upon averaging just 3-4 time series together, in most cases we observe strong correlations with dial data.

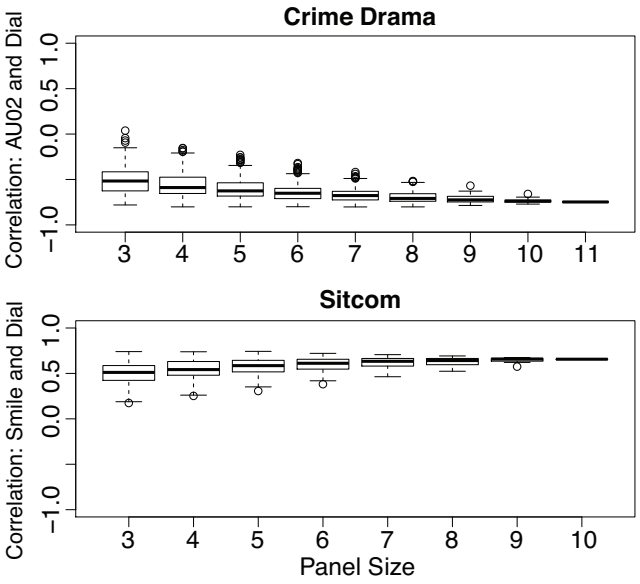

Fig. 4. Box plots assessing the lower limits of panel size.

VII. IDENTIFYING DIVERGENCE FROM THE DIAL

The correlations presented in Table III suggest that in a majority of cases, expressions are congruent with the dial responses. However, the dial data should not necessarily be treated as ground truth, as there may be significant caveats with this approach as described in the introduction. The utility of this approach may reside in the opportunity to explore temporal slices of substantial disagreement between dial data and facial expression data and subsequently mine those regions of facial video to search for deeper ancillary

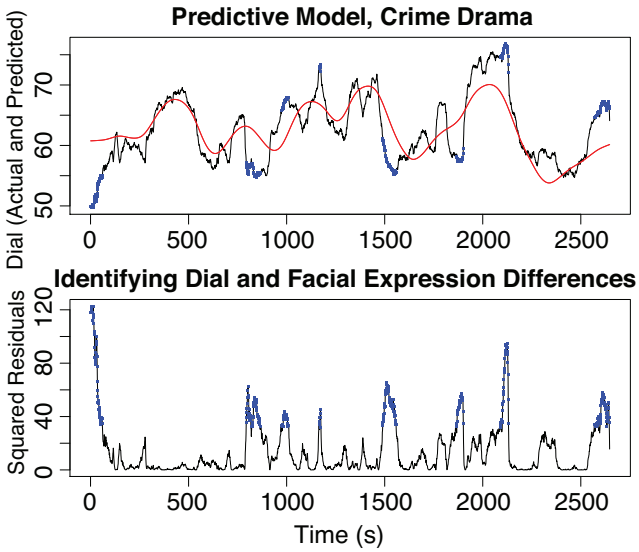

Fig. 5. Identifying regions of disagreement for the crime drama. Top) dial values (black) and model predicted values (red), bottom) squared residual differences between the dial and the predicted values. Blue highlights distinguish regions of greatest disagreement.

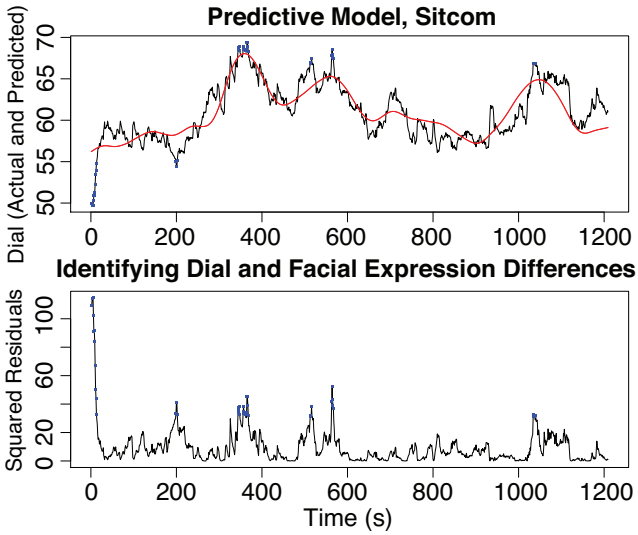

Fig. 6. Identifying regions of disagreement for the sitcom. Top) dial values (black) and model predicted values (red), bottom) squared residual differences between the dial and the predicted values. Blue highlights distinguish regions of greatest disagreement.

insights. In particular, this may serve to highlight moments when self-report measures are undermined. For example, panelists may experience high cognitive load when using the dial and it may not reflect their true latent affective state accurately in some temporal regions.

We present a method for quantitatively identifying moments where facial expressions differ from the dial: a predictive model, where the target variable y is the dial time series and the candidate predictor matrix **X** is composed of all AUs and V. From this predictive model, sufficiently large residuals can be extracted and can serve to characterize temporal locations where the facial expression data mismatches the dial. The intuition is that large residuals may serve to characterize regions where facial expression data may represent a fundamentally different latent emotional state of interest for deeper analysis.

We emphasize that a variety of predictive models could be constructed to predict y, such as but not limited to linear regression, generalized linear models, nonlinear regression, principal component regression, and support vector machines. In this paper, we chose mixed stepwise least absolute shrinkage and selection operator (LASSO) regression [22] as a predictive model. LASSO attempts to balance predictive accuracy (minimizing sum of squared errors) with model simplicity (bounded sum on absolute values of coefficients) by utilizing an L1-norm constraint that penalizes complexity while retaining straightforward coefficient interpretation similar to the popular least squares approach.

Letting **X**=[$AU02^T$ $AU04^T$ $AU09^T$ $Smile/AU12^T$ $Uni-AU12/14^T$ V^T] (where italics indicate a spline-smoothed time series vector and T denotes transposed), a forward stepwise LASSO model is estimated for the Crime Drama and the Sitcom. Specifically, the model begins with no terms, subsequently adding terms until the model is saturated with all available terms. Diagnostics, including residual sum of squares and Mallows criterion (Cp), are available with each additional term. We opted to strike a balance between number of predictors and accuracy, thus searching for an apparent minimization asymptotic for both criteria.

For the Crime Drama, error minimization appears to asymptote with the inclusion of just AU02, yielding an R^2 of 0.56. Subsequently, fitted values and squared residuals are plotted in Figure 5 to highlight regions that may be of interest for deeper investigation, as they imply disagreement between facial expressions and self-report. To highlight these differences, points on the dial where the squared residuals exceeds a chosen threshold are demarcated in blue. Specifically, the points are chosen by (i) calculating the mean and standard deviation from the lowest 40% order statistics of the squared residuals and (ii) enumerating squared residuals that exceed this mean plus 12 standard deviations.

The same procedure is conducted for the Sitcom (Figure 6). In this case, error minimization appears to asymptotes with the inclusion of Smile/AU12, AU9/AU10 and AU02 with an R^2 of 0.74. The Sitcom model fit is significantly better than that of the Crime Drama. One possible explanation could be that the Crime Drama show was twice as long as the sitcom, and may have resulted in fatigue of the dial task, or simply forgetting more often to turn the dial. Also, the Sitcom predictive model has three terms while the Crime Drama has only one, which explains a portion of the accuracy differential. The largest discrepancy in both cases occurs during the onset of the content. Panelists were told to start with the dial at 50 (neutral). However, the facial information correctly identifies the initial state as non-neutral, suggesting higher onset accuracy.

VIII. Discussion

AU02 and AU4 were particularly strong predictors of interest in the case of a crime drama and Smile was the

strongest predictor of interest in a sitcom, which is consistent with suspense and amusement, respectively, being key indicators of interest for these two genres. This type of insight would not be possible with a 1-dimensional dial recording. By measuring multiple facets of facial response we gain a depth in understanding of emotional response. As another example, in media content, such as advertising, there is common use of violation of expectation (surprise) as a way of creating humor [1]. With a 1-dimensional dial measurement, the distinction between these two states will not necessarily be identifiable or quantifiable, but measuring multiple facets of facial activity yields insight about the interactions between different states.

Using a predictive model such as the one presented, provides a systematic means for exploring temporal regions of difference between dial and facial expression data. For example, in the Crime Drama, predicted values seem to diverge from actual dial data in cases where panelists are "shocked" by a moment in the program; when this occurs, there is a delay in the dial response that is not observed in the face. For the sitcom, there is a dissonance between moments where the laugh track (an artificial sound byte intended to encourage real viewers to laugh) is active but panelists are not smiling. Although it is unclear whether the dial or facial expression data is a better indicator of the latent emotional state, it is reasonable to assume that facial expressions might be a better reflection of the person's true state.

IX. Conclusions and Future Work

This paper presents the first systematic evaluation of self-reported dial data compared to data from an automated facial analysis system. Our system detects a range of facial action units as well as a continuous measure of valence in spontaneous facial videos. Results suggest that, automated facial expression analysis can serve as an accurate proxy for the self-report dial method of measurement and potentially yields deep insights beyond it. We presented a method for finding the moments where self-report and facial expression data diverge, providing additional insights over self-report alone. We also present an exploration of the sample size needed for media research studies and show that facial analysis requires a panel size that is 5% of the size needed for dial studies. The results also show that the probability outputs predicted by the classifiers correlate with the intensity of the underlying state, in this case reported interest. However, further studies in which the participants are asked to report valence and arousal specifically, rather than interest, would help strengthen this finding.

More experiments are needed to develop an understanding of what specific facial expressions are related to dial data for various media genres. For the two genres explored in this work (comedy and drama), additional trials with similar programs will be needed to test the robustness and generalizability of the specific correlations found here. Standard emotion eliciting movie sequences such as those identified by Gross et al. [5] could also be tested for further validation.

Acknowledgments

We would like to thank Zhihong Zeng, Jay Turcot, Daniel Bender, Melissa Burke and Andy Dreisch for contributions to this work. We would like to thank the partners who helped collect the data and the panelists who took part in the study.

References

[1] D. Alden, A. Mukherjee, and W. Hoyer. The effects of incongruity, surprise and positive moderators on perceived humor in television advertising. *Journal of Advertising*, pages 1–15, 2000.

[2] J. Cohn, Z. Ambadar, and P. Ekman. *Observer-based measurement of facial expression with the Facial Action Coding System*. Oxford: NY, 2005.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. Ieee, 2005.

[4] P. Ekman and W. Friesen. Facial action coding system. 1977.

[5] J. Gross and R. Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.

[6] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99, 2010.

[7] R. Hazlett and S. Hazlett. Emotional response to television commercials: Facial emg vs. self-report. *Journal of Advertising Research*, 39:7–24, 1999.

[8] C. Hjortsjö. *Man's face and mimic language*. Studen litteratur, 1969.

[9] M. Lieberman, N. Eisenberger, M. Crockett, S. Tom, J. Pfeifer, and B. Way. Putting feelings into words. *Psychological Science*, 18(5):421, 2007.

[10] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.

[11] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[12] D. McDuff, R. El Kaliouby, K. Kassam, and R. Picard. Affect valence inference from facial action unit spectrograms. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 17–24. IEEE.

[13] D. McDuff, R. El Kaliouby, and R. Picard. Crowdsourced data collection of facial responses. In *Proceedings of the 13th international conference on Multimodal Interaction 2011*. ACM.

[14] D. McDuff, R. El Kaliouby, and R. Picard. Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing*, 3(4):456–468, 2012.

[15] G. Mckeown, M. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Proceedings of IEEE Int'l Conf. Multimedia, Expo*, pages 1079–1084, July 2010.

[16] M. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, 2(2):92–105, 2011.

[17] A. Ruef and R. Levenson. Continuous measurement of emotion. *Handbook of emotion elicitation and assessment*, pages 286–297, 2007.

[18] D. Sander and K. Scherer. *Oxford companion to emotion and the affective sciences*. Oxford University Press, USA, 2009.

[19] B. Schuller, M. Valstar, R. Cowie, and M. Pantic. Avec 2012–the continuous audio/visual emotion challenge. In *Proceedings 2nd International Audio/Visual Emotion Challenge and Workshop, AVEC*, 2012.

[20] T. Senechal, J. Turcot, and R. El Kaliouby. Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience. In *Automatic Face and Gesture Recognition, 2013. Proceedings. Tenth IEEE International Conference on*, 2013.

[21] T. Teixeira, M. Wedel, and R. Pieters. Emotion-induced engagement in internet video ads. *Journal of Marketing Research*, 2010.

[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[23] F. Torre and J. Cohn. Facial expression analysis. *Visual Analysis of Humans*, pages 377–409, 2011.