2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)

DynAMoS: The Dynamic Affective Movie Clip Database for Subjectivity Analysis

Jeffrey M. Girard
Department of Psychology
University of Kansas
Lawrence, KS, USA
0000-0002-7359-3746

Yanmei Tie
Brigham and Women's Hospital
Harvard Medical School
Boston, MA, USA
0000-0001-8345-8903

Einat Liebenthal
McLean Hospital
Harvard Medical School
Belmont, MA, USA
0000-0003-4917-8160

***Abstract*—In this paper, we describe the design, collection, and validation of a new video database that includes holistic and dynamic emotion ratings from 83 participants watching 22 affective movie clips. In contrast to previous work in Affective Computing, which pursued a single "ground truth" label for the affective content of each moment of each video (e.g., by averaging the ratings of 2 to 7 trained participants), we embrace the subjectivity inherent to emotional experiences and provide the full distribution of all participants' ratings (with an average of 76.7 raters per video). We argue that this choice represents a paradigm shift with the potential to unlock new research directions, generate new hypotheses, and inspire novel methods in the Affective Computing community. We also describe several interdisciplinary use cases for the database: to provide dynamic norms for emotion elicitation studies (e.g., in psychology, medicine, and neuroscience), to train and test affective content analysis algorithms (e.g., for dynamic emotion recognition, video summarization, and movie recommendation), and to study subjectivity in emotional reactions (e.g., to identify moments of emotional ambiguity or ambivalence within movies, identify predictors of subjectivity, and develop personalized affective content analysis algorithms). The database is made freely available to researchers for noncommercial use at `https://dynamos.mgb.org`.**

***Index Terms*—database, emotion elicitation, content analysis, affective computing, subjectivity, multimodal, movie clips**

I. Introduction

In the affective and medical sciences, it is common for researchers to use standardized stimuli such as text vignettes, images, audio clips, and video clips to (try to) elicit emotions in their participants [1]. This methodology can be used to study the effects of emotion on various psychological and physiological processes, to identify correlates of and group differences in emotional reactivity, and even to detect and quantify affective dysfunction in individual participants. The focus in such studies is on the affective experiences of the *participants* themselves (i.e., what they feel in response to the stimuli) as opposed to their perceptions of others (e.g., what the people portrayed in the stimuli seem to be feeling).

Movie clips are popular stimuli in such studies because movies are often expertly crafted (e.g., by actors, directors, sound designers, and editors) to produce a wide range of emotional reactions in viewers. Due to their multimodal nature and extended duration, they tend to elicit stronger and more complex emotional responses than other stimulus types [2].

Participants are typically asked to provide a single, *holistic* report of their emotional reaction after watching each movie clip (e.g., in terms of *discrete* categories like anger and sadness or *continuous* dimensions like valence and arousal [3]). However, participants' emotions often evolve over time during a movie clip. To address this issue, specialized methods have been developed to collect *dynamic* reports during stimulus presentation (e.g., by having the participant move a dial or lever to indicate changes in emotion) [4]. The resulting time-series (sometimes called 'traces') can capture the moment-to-moment unfolding of emotional reactions over time with high granularity [5].

In the Affective Computing community, dynamic ratings of experienced and perceived emotion have been used to create 'ground truth' labels for emotion recognition and affective content analysis [3]. However, because each rater is a unique individual with their own history, background, and constellation of affective traits, there is inevitably some degree of inter-rater variability or 'subjectivity' in their ratings of each stimulus. This variability is often considered a nuisance—a source of noise to be minimized, e.g., by averaging across raters, training raters to consensus, or switching from continuous rating scales to ordinal rankings [6]. However, we contend that inter-rater subjectivity is actually a fascinating phenomenon worthy of academic study in its own right.

Embracing the existence of this subjectivity leads to many intriguing questions. Why do different participants experience the same stimuli in such different ways? How structured and predictable are individual participant's responses? Are some stimuli (or parts of stimuli) associated with greater inter-participant variability than others, and if so, how well can we estimate their degree of subjectivity based on their content alone? We believe that the Affective Computing community is well-suited to begin answering these questions; however, doing so will require new datasets and novel methods.

In service of this goal, we present the DynAMoS database, which we designed to facilitate research on the dynamic and subjective aspects of emotional reactions to movie clips.

This research was partially supported by grant R01 DC020965 from the National Institutes of Health (NIH).

A. Previous Databases

In order to position our database in the broader literature and highlight the novelty of our approach, we briefly review previous databases (sometimes also called 'stimulus sets') that contain emotion ratings of affective movie clips.

Dozens of current databases contain affective movie clips (see [7] for a recent review). However, the vast majority of such databases only include holistic emotion ratings and not dynamic emotion ratings; this omission precludes the analysis of within-stimulus changes and temporal patterns of reactivity. Furthermore, most databases that do include dynamic ratings (e.g., [8], [9]) are focused on participants' perceptions of others' emotions and not participants' reports of their own emotional reactions. Emotion perceptions are also fascinating and worth studying, but this is a fundamentally different research question [10]. There are also several databases that include dynamic emotion ratings of other types of videos (e.g., [11]–[13]) but these also mostly focus on perceived emotion.

To our knowledge, there are currently only two databases that include dynamic ratings of participants' self-reported emotional reactions to affective movie clips. DECAF [14] had 7 participants watch 36 short (i.e., 1–2 minute) clips and dynamically rate their own emotional valence and arousal at 1 Hz. Similarly, COGNIMUSE [15] had 7 participants watch 7 long (i.e., 30 minute) clips and dynamically rate their own emotional valence and arousal at 25 Hz. Although a somewhat different context than movie clips, there is also the TVNEWS [16] database that had 50 participants watch 144 short (i.e., 20–52 second) clips from television news programs and dynamically rate their own emotional valence at 2 Hz.

Thus, there are only a few video databases that include dynamic ratings of self-reported emotion. These databases also have several characteristics that limit their usefulness for subjectivity analyses and personalized modeling. Ideally, a large and representative sample of participants would dynamically rate each clip and the ratings from each participant would be separately provided with the database. However, DECAF and COGNIMUSE have small convenience samples ($n = 7$), which limit their generalizability and the statistical power of subjectivity analyses, and TVNEWS only provides the average rating (across all participants) at each moment, which makes it unusable for subjectivity and personalized modeling. Additionally, while longer clips contain more contextual information and can yield more participant engagement/investment, they are also more likely to fatigue participants and be emotionally heterogeneous [17]. Thus, the ideal clip length is somewhere around 2–10 minutes, which is a bit longer than the clips in TVNEWS and a lot shorter than those in COGNIMUSE.

We contend that the affective sciences need a new database focused on subjectivity and personalized modeling of emotional reactions to affective movie clips. To meet this goal, it should recruit a large number of diverse participants to dynamically rate each clip, use clips that are 2–10 minutes in duration, and provide each participant's ratings separately.

B. The Current Paper

The current paper presents a new database that meets this need and begins to fill in the gaps of previous databases. We had 83 participants watch 22 medium-length (i.e., 2–7 minute) movie clips and dynamically rate their own emotional valence (and holistically rate their positive and negative affect). This number of raters is an order of magnitude greater than the number of raters in the DECAF and COGNIMUSE databases and, unlike the TVNEWS database, we provide every participant's individual ratings. As a result, the DynAMoS database opens up exciting new possibilities into the study of affective dynamics, subjectivity, and personalized modeling.

Due to space constraints, we focus the current paper on describing the database and validating its measures rather than providing baseline models or case studies. We recognize that this focus is uncommon in the ACII community but argue that measurement validation is often underappreciated and worth the added emphasis here. An extended version that includes feature descriptions, cross-validation partitions, and baseline results will be submitted as a separate full-length paper.

The contributions of the current paper are five-fold:

1) We propose a novel approach to dealing with and thinking about inter-rater variability (i.e., 'subjectivity'): to embrace it as stemming from individual differences that are worth studying and explicitly modeling rather than a nuisance to be merely corrected or mitigated.
2) We propose a method for collecting data to study inter-rater variability, which adapts existing methods but meaningfully changes their goals and priorities.
3) We provide a new database of affective movie clips with dynamic and holistic ratings of experienced emotion from a large and diverse sample of participants and discuss new applications for this type of data.
4) We demonstrate the use of modern measurement validation techniques including Bayesian generalizability studies, inter-rater reliability analysis with incomplete data, and coefficient categorical omega for estimating the internal consistency of ordinal scale scores.
5) We develop a novel data visualization approach for depicting inter-rater variability in time series data.

II. Methods

A. Movie Clip Selection

Clips from feature films of different genres (e.g., comedy, romance, drama, action) were identified using open libraries. Selection of the final set of clips was based on the following criteria: (1) each clip must be 2–10 minutes in duration, (2) each clip must contain dialogue spoken in English by live-action human actors in largely camera-facing orientations, (3) the set as a whole must evoke a range of emotional reactions spanning positive and negative valence of various intensity levels, and (4) the set as a whole must represent a diversity of actor demographics (e.g., age, sex, and race).

B. Movie Clip Processing

Each clip was extracted from a Blu-Ray copy of its source movie to a separate MPEG-4 file with a frame width of $1920\,\mathrm{px}$, a frame rate of $23.976\,\mathrm{fps}$, and an audio sampling rate of $48\,\mathrm{kHz}$. The official English-language audio track was used. (English-language subtitles were extracted for later analysis but were not shown to participants in the experiment.)

C. Participant Recruitment

Participants were recruited from the community using the Rally with Mass General Brigham (MGB) online platform; they were all living in the USA at the time of participation. Applicants were first screened to confirm their eligibility for the study; inclusion criteria included (1) having no uncorrected sensory, cognitive, or emotional impairments, (2) being age 18–60 years old, (3) being fluent in English, and (4) having access to a laptop or desktop computer and a quiet environment for the study sessions. After signing a consent form, participants were asked about their demographic background (i.e., age, sex/gender, and race/ethnicity).

D. General Procedure

Participation occurred remotely via video conferencing (with the experimenter's camera and microphone turned off during video-watching and rating). To prevent fatigue, participants completed the experiment across two 90-minute sessions. In each session, they independently viewed and rated 11 movie clips presented in randomized order. Before each movie clip, a brief and standardized description was read aloud by the experimenter. These descriptions were meant to orient participants to the scene and provide any necessary contextual information without describing the emotional tone (see the website for all descriptions). The participant then watched the clip while simultaneously providing dynamic valence ratings (as described below). After the clip ended, the participant provided holistic emotion ratings (as described below) and answered several other questions about whether the audio and video playback worked properly and how familiar they were with the clip's source movie. After each session, participants were compensated 25 USD for their effort; thus, each participant could earn up to 50 USD in total. This compensation rate is about a dollar higher than the current minimum hourly wage in the state of Massachusetts.

E. Affect Rating Procedure

Participants watched each clip and simultaneously provided dynamic valence ratings using the CARMA software [18]. As depicted in Figure 1, CARMA displays the video next to a vertical rating scale, represented by a color gradient, that ranged from -4 (*very negative*) to $+4$ (*very positive*). Participants were instructed to move a slider up and down within this rating scale (using their mouse or arrow keys) to reflect how negative/unpleasant to positive/pleasant the movie clip made them feel from moment to moment. The slider was positioned at zero at the onset of each clip, and participants could adjust the slider at any time; CARMA queried the relative location of the slider at $30\,\mathrm{Hz}$ and then averaged all queried values within $1\,\mathrm{s}$ temporal bins. Our rationale for this aggregation step is that it smooths the time series, removing unintentional motion artifacts, and yields scores that better align with the response time needed to process each moment of the clip (e.g., sensorially, emotionally, and cognitively) and make an appropriate motor response; it also helps align the ratings of participants who may differ in their reaction times. Although approaches exist for trying to align participants' ratings using complex algorithms or additional data collection [19], we prefer the simplicity of the aggregating approach.

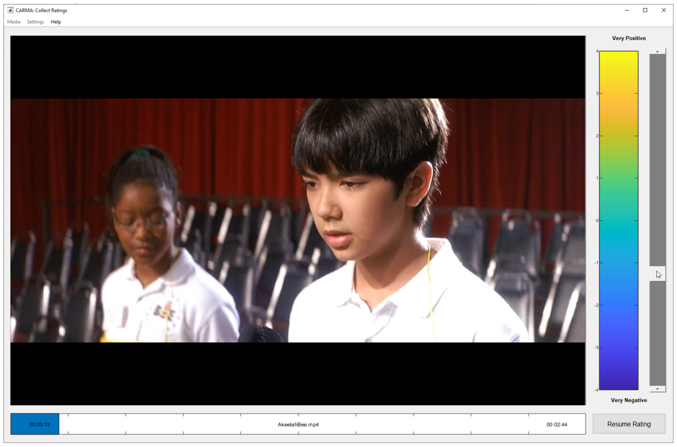

Fig. 1. Screenshot of Continuous Rating Collection in CARMA [18]

After watching each clip, participants provided holistic affect ratings using the Short Positive and Negative Affect Scale (S-PANAS) [20], [21]. Participants were instructed to think about their overall emotional reaction to the movie clip and rate it on five positive affect items (alert, determined, enthusiastic, excited, inspired) and five negative affect items (afraid, distressed, nervous, scared, upset) using ordinal scales from 0 (*very slightly or not at all*) to 4 (*extremely*). Scores on these items were combined through averaging to yield scale scores for positive affect and negative affect.

F. Validation Procedure

We first excluded the ratings of participants who reported that a clip's audio or video did not play properly. Given our interest in inter-rater variability, we chose not to exclude participants for being outliers in terms of their ratings.

We then estimated inter-rater reliability of the valence ratings within each movie clip (after excluding the first $10\,\mathrm{s}$ of each clip for reasons described in §III-D) and of the holistic ratings across all movie clips. Specifically, two-way intraclass correlation coefficients (ICCs) were estimated from Bayesian generalizability studies [22] using the `varde` R package [23].[1] There are many formulations of the two-way ICC, but the most relevant here are the single-measures consistency ICC for

[1] Note that generalizability studies assume uncorrelated facet levels, which is violated by the autocorrelation of adjacent bins. However, a simulation study [24] found that the bias caused by autocorrelation decreases as the number of bins increases; thus, with 130–425 bins per video, this bias is mitigated.

incomplete data or ICC($Q, 1$), which quantifies the reliability of the ratings from a single randomly selected rater [22] and the average-measures consistency ICC for incomplete data or ICC($Q, \hat{k}$), which quantifies the reliability of the average of all available raters' ratings. We followed common heuristics [25] in considering ICC values above .90 to be excellent.

From these same generalizability studies, we also calculated the percentage of rating variance that was accounted for by the differences between rater intercepts; this percentage can be considered a rough index of how much rater-to-rater subjectivity was present in the ratings for each clip.

We also estimated the inter-item reliability (or internal consistency) of the holistic scales for each movie clip. To do so, we estimated coefficient categorical omega or $\omega_{u\text{-}cat}$ [26], [27] for the Positive Affect and Negative Affect scales from ordinal confirmatory factor analysis models of each clip's holistic ratings (using the `lavaan` [28] and `semTools` [29] R packages). We followed common heuristics [30] in considering omega values above .75 to be acceptable.

G. Website Generation

As a form of rich documentation for the database, we used the Quarto technical and scientific publishing system[2] to create a website for the database within R. This website reads in the database files and generates summary statistics, tables, and figures for the database as a whole and for each movie clip individually. It also includes screenshots from movie clips, word clouds of the subtitles, and visualizations of the ratings. These pages are all parameterized reports, which means they can be quickly and easily updated as the database grows and changes. The website is hosted using GitHub Pages.[3]

III. Results

A. Movie Clip Summary

Our final set includes 22 movie clips, each drawn from a different English-language feature film. As shown in Table I, the source movies were released between the years of 1991 and 2018 and the clips ranged from 130–425 seconds in duration (M=251.1, SD=90.8). Example video frames from the movie clips are shown in Figure 2; even from this small sample of images, it is possible to see the diversity of characters, settings, and lighting conditions represented in the database.

B. Participant Summary

We recruited a total of 83 participants. In terms of sex/gender, 56 reported being Female (68 %) and 26 reported being Male (31 %). In terms of race, 43 reported being White (52 %), 22 reported being Asian (27 %), 12 reported being Black (15 %), and 5 reported being another race (6 %). In terms of ethnicity, 70 reported being non-Hispanic/Latino (84 %) and 11 reported being Hispanic/Latino (13 %). In terms of age, participants ranged 18–59 years old (M=28.8, SD=9.9).

C. Validation Results

Of the 83 recruited participants, 77 (93 %) completed both sessions and the remaining 6 only completed the first session. Out of the 1826 possible participant-clip combinations, data from 1702 (93 %) were collected without issue, data from 93 (5 %) were not collected (due to participant dropout and clips being added partway through recruitment), and data from 31 (2 %) were excluded due to participants reporting issues with audio and/or video playback.

Estimates of the inter-rater reliability of the dynamic valence ratings for each movie clip are presented in Table I (and more detailed results are provided on the database website). The reliability of a single, randomly selected rater was quite poor (i.e., ICC$(Q, 1) < 0.50$) for all clips except three. We thus cannot be very confident that any single rater's scores will represent the rating of a "typical" participant. This result speaks to the subjectivity inherent to the ratings and helps motivate its study. Similarly, rater intercepts explained a substantial amount of variance in the ratings, ranging from 21 % to 65 % (also shown in Table I), which implies considerable subjectivity in the continuous ratings. However, the reliability of the average of all available ratings per bin was "excellent" (i.e., ICC$(Q, \hat{k}) > 0.90$) across all clips. These results imply that, for applications that use the average rating per bin, it can be used with high confidence in all clips to represent the rating of a "typical" participant. These two sets of results may at first seem to be contradictory but are in fact expected, as the latter is largely due to the fact that averaging many raters offsets the idiosyncratic aspects of individual ratings [31].

Estimates of the inter-rater reliability of the holistic ratings cannot be calculated per movie clip (since they are only provided once per clip) and instead must be calculated across all clips. The reliability of the average of all available ratings was "excellent" (i.e., ICC$(Q, \hat{k}) > 0.90$) for all items as well as for the Positive Affect scale (ICC=.990) and the Negative Affect scale (ICC=.988) scores. (Item-level results and interval estimates are provided on the database website.) These results imply that the average rating per clip can be used with high confidence to represent the rating of a "typical" participant. The amount of variance in the holistic ratings explained by rater intercepts was 24.5 % for the Positive Affect scale and 19.3 % for the Negative Affect scale. These results imply that there was relatively less subjectivity in the holistic ratings as compared to the dynamic valence ratings.

Estimates of the inter-item reliability of the holistic ratings (within raters) can be calculated per movie clip (since each rater provides five ratings per scale). The categorical omega estimates per movie clip ranged from .66 to .89 (M=.81, SD=.06) for the Positive Affect scale and from .61 to .92 (M=.82, SD=.09) for the Negative Affect scale. Thus, most (but not all) clips had acceptable internal consistency when using the holistic ratings of any single participant.

D. Dynamic Rating Summary

The distribution of dynamic valence ratings across all raters, bins, and movie clips is depicted in Figure 3. This distribution

[2] https://www.quarto.org

[3] https://pages.github.com

TABLE I

Summary Information About the Movie Clips, Source Movies, Holistic Ratings, and Continuous Ratings

			Source Movie Information		Holistic PA		Holistic NA		Dynamic Valence			
Clip	Duration	Raters	Title	Year	Mean	Omega	Mean	Omega	Mean	ICC$_{Q1}$	ICC$_{Qk}$	Rater
01	164s	81	Akeelah and the Bee	2006	2.49	.84	0.35	.78	1.51	.698	.995	23.3%
02	162s	76	If Beale Street Could Talk	2018	1.19	.84	0.66	.86	0.57	.192	.950	39.3%
03	274s	79	Catch Me If You Can	2002	1.16	.76	0.45	.85	0.30	.408	.982	22.4%
04	132s	80	500 Days of Summer	2009	0.81	.78	0.42	.86	0.21	.180	.947	50.7%
05	134s	79	Fences	2016	0.78	.72	1.79	.91	−2.06	.239	.963	52.5%
06	218s	78	Forrest Gump	1994	0.96	.74	0.59	.78	−0.49	.145	.932	47.9%
07	150s	76	Good Will Hunting	1997	0.92	.81	0.93	.88	0.16	.398	.981	41.7%
08	239s	75	The Green Mile	1999	0.73	.78	2.24	.89	−2.57	.204	.951	45.5%
09	425s	77	The King's Speech	2010	1.69	.88	0.80	.86	0.45	.185	.946	49.0%
10	232s	73	Lady Bird	2017	0.99	.87	0.36	.71	0.18	.335	.974	25.2%
11	414s	75	Legally Blonde	2001	2.20	.83	0.24	.72	1.15	.319	.973	41.0%
12	130s	75	Little Miss Sunshine	2006	0.70	.66	0.46	.74	−0.45	.113	.910	70.4%
13	240s	76	Marriage Story	2019	0.73	.88	1.49	.84	−2.18	.264	.965	58.8%
14	341s	78	Miracle	2004	2.19	.89	0.31	.72	1.10	.686	.994	21.1%
15	234s	75	Moonlight	2016	1.59	.86	0.47	.88	1.19	.291	.969	46.2%
16	262s	72	No Country for Old Men	2007	0.83	.84	1.23	.92	−0.99	.224	.954	58.2%
17	425s	78	The Parent Trap	1998	1.49	.84	0.19	.62	1.90	.279	.968	38.7%
18	229s	79	Pulp Fiction	1994	0.95	.85	0.81	.91	0.21	.172	.943	59.5%
19	237s	79	The Pursuit of Happyness	2006	2.01	.87	0.33	.68	1.35	.495	.987	32.9%
20	274s	79	The Silence of the Lambs	1991	0.84	.72	1.57	.90	−1.37	.268	.967	47.8%
21	294s	73	The Social Network	2010	0.91	.81	0.54	.79	−0.50	.140	.923	40.7%
22	315s	75	Zodiac	2007	0.97	.79	1.85	.91	−0.99	.628	.992	29.5%

Note. PA and NA = Positive and Negative Affect, Omega = Internal Consistency, ICC = Intraclass Correlation, Rater = Rater Variance Percent.

Fig. 2. One Example Video Frame from Each Movie Clip (arranged left-to-right in the same order presented in Table I)

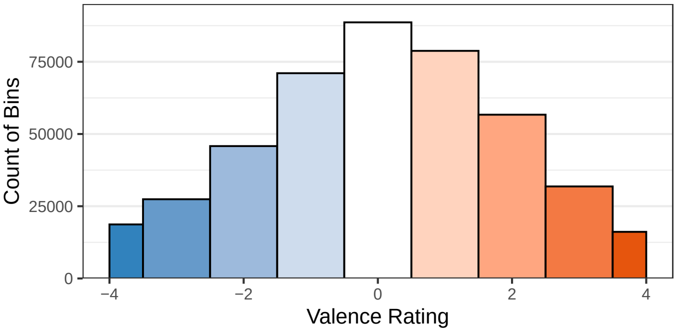

Fig. 3. Histogram of All Dynamic Valence Ratings

has a roughly Gaussian shape with ratings close to 0 being the most common and increasingly extreme ratings (in either direction) being increasingly less common.

Figure 4 displays the time series of dynamic valence ratings from four example movie clips (similar plots for all movie clips are available on the database website). The thick black line on each plot depicts the average of all available ratings per temporal bin; this is the highly reliable time series that users can use to represent the rating of a "typical" participant. The colored ribbons around the black line depict successively larger percentages of the ratings per bin, i.e., the yellow ribbon contains the most common ratings, the green bands show less common ratings, and the purple band shows even less common ratings. This novel visualization approach, which we call a "chromodoris plot" (after the colorful sea slugs of similar appearance), allows us to quickly see the central tendency of ratings as well as their spread (e.g., to locate parts of each movie clip that were more or less subjective).

A few insights can be derived from these visualizations. First, the ratings of the first 5 s to 10 s of each movie clip typically centered around zero with minimal spread. This pattern is likely due to the raters getting oriented to each clip. For many applications of the database (e.g., predicting ratings' mean or spread), it would make sense to exclude these bins from analysis. Second, some movie clips (e.g., *Akeelah and the Bee*) were relatively stable (or "stationary") in terms of their spread in ratings, whereas other clips had moments of sharp deviation. A striking example of the latter case is from *The Green Mile*; the ratings are quite negative throughout this clip, but at two moments (i.e., 01:00 to 01:30 and 3:00 to 3:40) the mean ratings became less negative and the spread in ratings increased dramatically. Similar increases in rating variability occur in both the *Fences* and *Lady Bird* time series, albeit to a lesser degree. The *Lady Bird* time series is especially

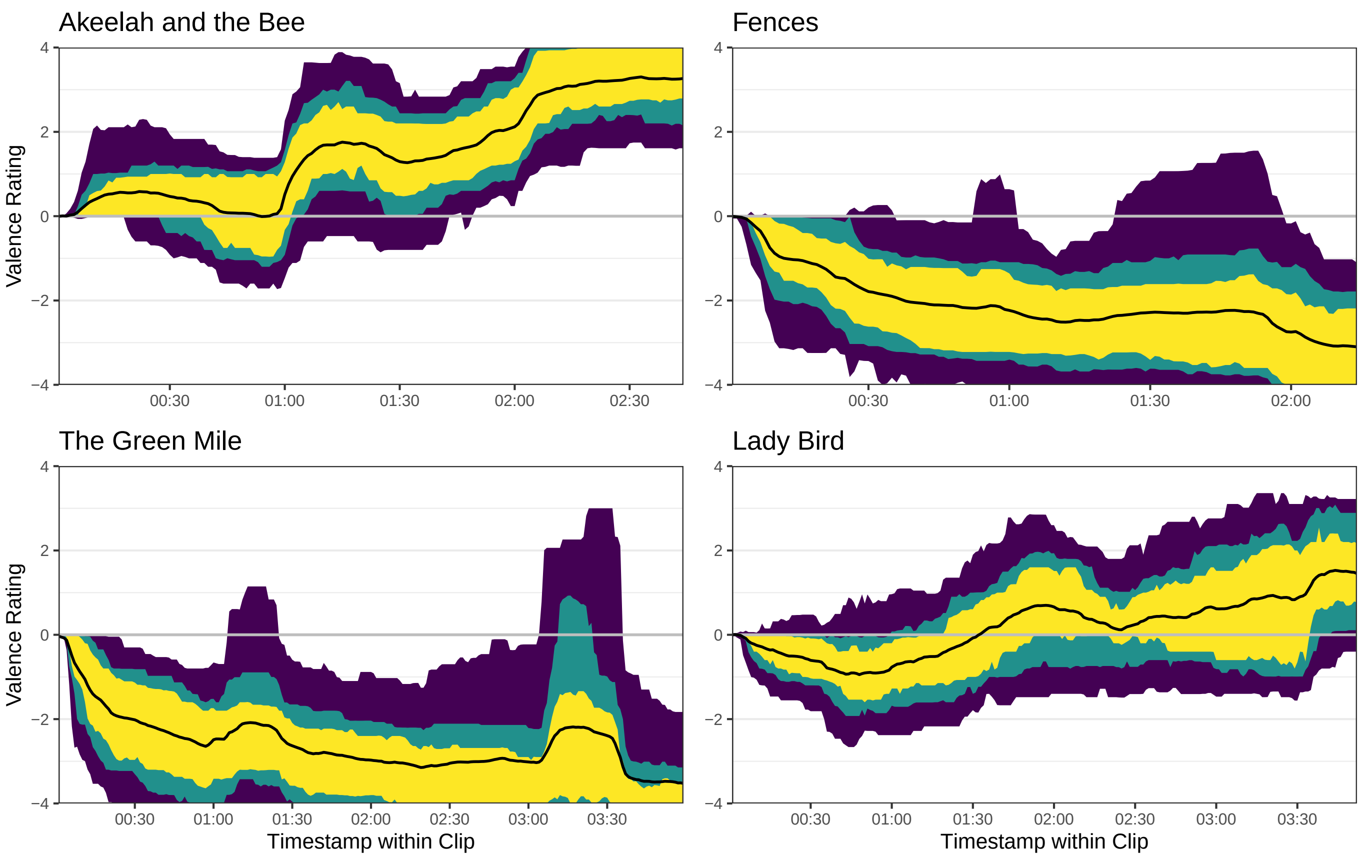

Fig. 4. Example Chromodoris Plots of Dynamic Ratings (Black = Mean Rating, Yellow = Inner 50%, Green = Inner 70%, Purple = Inner 90%)

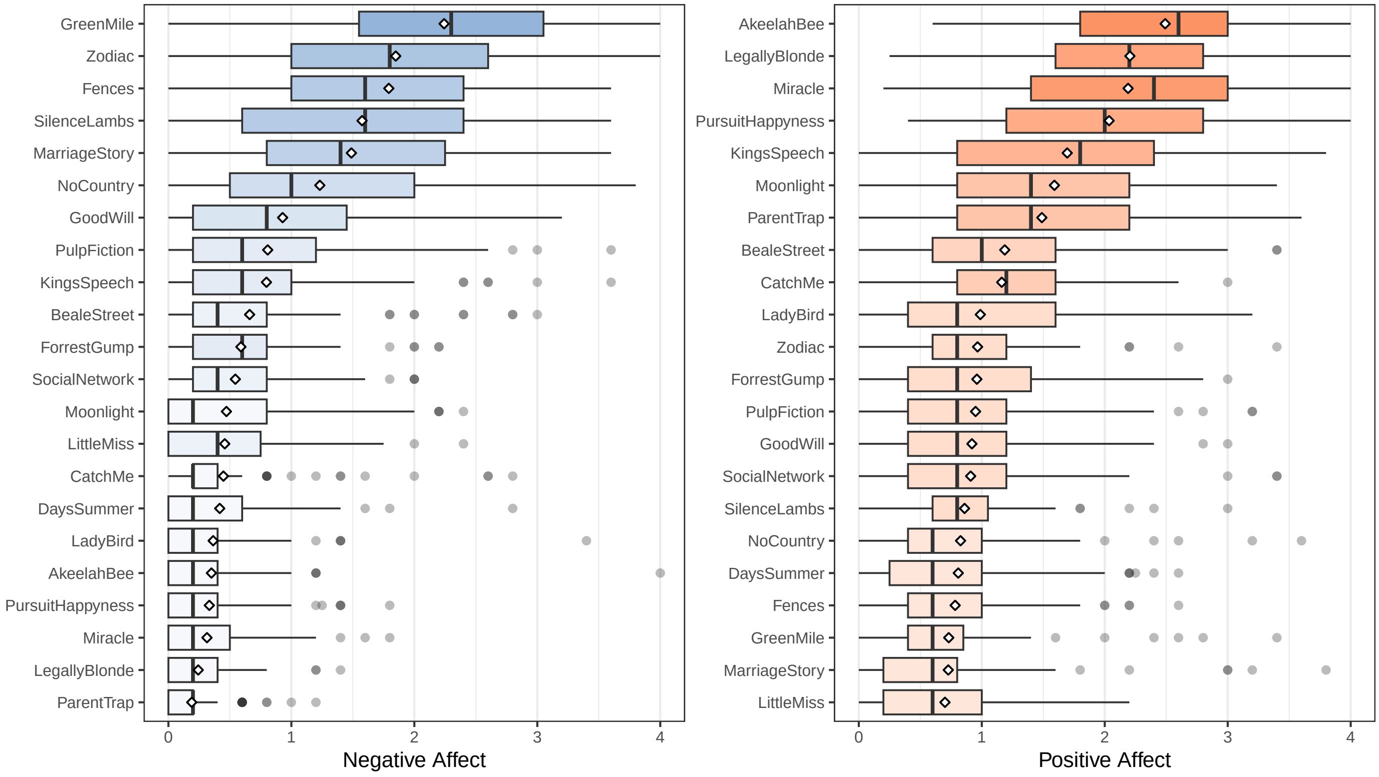

Fig. 5. Boxplots of the Distribution of Holistic Affect Ratings for Each Movie Clip (white diamonds = means, grey circles = outliers)

interesting because large portions of the raters disagreed at many points whether the clip was positive or negative. These patterns raise several questions: What is it about these specific moments that leads to increased variability? Can we predict such moments from their content? What is it about the raters that caused them to respond differently to these moments? Can we predict such deviations from the norm?

E. Holistic Rating Summary

The distribution of holistic affect ratings for each movie clip is depicted in Figure 5. Clip averages ranged from 0.70 to 2.49 for Positive Affect (M=1.23, SD=0.56) and from 0.19 to 2.25 for Negative Affect (M=0.82, SD=0.60). Four clips exceeded 2.0 and nine clips exceeded 1.0 for average Positive Affect. In contrast, only one clip exceeded 2.0 and only six clips exceeded 1.0 for average Negative Affect.

IV. Discussion

We argue that subjectivity (i.e., inter-rater variability) is inherent to emotion representation and encourage researchers in Affective Computing to study this phenomenon rather than simply try to control for it. We argue that doing so has the potential to unlock new research directions, generate new hypotheses, and inspire novel methods. To promote work on this topic, we created and are sharing a new database with dynamic and holistic ratings of dozens of participants' emotional reactions to movie clips. In this paper, we describe the design, collection, and validation of the database. Results from our validation analyses support the trustworthiness of the database and reveal a large degree of subjectivity to be analyzed. We hope that this database will prove useful to researchers in several disciplines and will inspire more work on this interesting phenomenon (and related ones like ambiguity in emotion perceptions [32]).

A. Database Uses

The DynAMoS database has many potential use cases across several disciplines. First, it can be used as a standardized set of videos for emotion elicitation with normative data on emotional reactions (both holistic and dynamic). For example, these movie clips could be shown to new participants in psychology, medical, and neuroscience studies to induce positive and negative affect; furthermore, each new participant's ratings could be compared to the distribution of ratings in the database to quantify deviations from the norm.

Second, it can be used to train and test affective content analysis algorithms using traditional methods. For example, the multimodal information contained in each movie clip (e.g., images, speech, music, and subtitles) could be used to predict the holistic and/or dynamic ratings *averaged across raters*. Such predictions could be helpful for video summarization, movie recommendation, and identifying moments-of-interest.

Third, it can be used to study subjectivity in affective experiences. For example, statistical or machine learning models could be used to predict the distributions of the holistic and/or dynamic ratings across raters. Possible predictor variables could include features of the movie clip (as in the second use case) and/or features of the participants themselves (e.g., demographics). Relatedly, this dataset may also be used in experiments on personalized/idiographic modeling [33] (e.g., predicting the ratings of specific individuals).

B. Database Access

Summary information about the database is available on the database website (see the link in the abstract) and access to the full database (i.e., video clips, video metadata, features, deidentified participant demographics, and dynamic and holistic ratings from each individual participant) will be granted free-of-charge to researchers for noncommercial use. Potential users will need to request access through a form on the database website and sign a licensing agreement.

C. Limitations and Future Directions

Limitations of the current study include: (1) a focus on English-speaking movies and participants living in the USA, which limits generalizability to other languages and populations, (2) a relatively small number of movie clips, which limits how varied our set can be, (3) a focus on valence and the positive/negative activation model [34], which does not exhaustively capture the affective domain, and (4) relatively little information was collected about each participant, which limits our ability to study the sources of individual differences.

To address these limitations in future work, we (1) invite collaborations with researchers from other countries, (2) plan to add more movie clips that cover additional combinations of emotional content, actor demographics, spoken languages, and recording conditions, (3) plan to collect holistic ratings of discrete emotions and appraisal dimensions, and (4) plan to collect additional self-report measures of relevant characteristics such as personality and mental health.

We could also collect dynamic ratings of additional affective dimensions (e.g., arousal), but the costs of doing so would be non-trivial as it would require participants to either repeat the rating procedure or simultaneously rate multiple dimensions (which is possible but challenging [35], [36]). Also, at least in the case of arousal, separate rating may not be fully necessary as prior research suggests that arousal tends to increase with the intensity of both positive and negative emotion (i.e., with the magnitude or absolute value of valence) [34], [37].

We plan to release an extended version of this paper and database that adds multimodal features extracted from the movie clips, standardized partitions for cross-validation, and baseline predictive models for the use cases described above.

Ethical Impact Statement

This work was approved by the governing Institutional Review Board prior to the work being carried out. Participants provided informed consent to complete the study and to have their deidentified data shared with other researchers. To further protect the security and confidentiality of the data, we require users of the database to sign a licensing agreement (§IV-B).

We believe that the risk of our work having negative societal impacts is low, especially if its limitations (§IV-C) are appreciated by readers and users. The main ethical question we wrestled with regarded our use of clips from copyrighted movies. We believe that our use of brief clips to induce emotions during noncommercial research constitutes "fair use" according to §107 of the U.S. Copyright Act and is highly unlikely to harm the market for the copyrighted works.

Acknowledgements

We thank Dr. Alexandra Golby for her support of this study, and Colin Galvin, Laura Rigolo, Abby Recko, and Ben Phan for their assistance with data collection.

References

[1] J. A. Coan and J. J. B. Allen, Eds., *Handbook of emotion elicitation and assessment*. Oxford University Press, 2007.

[2] R. Westermann, K. Spies, G. Stahl, and F. W. Hesse, "Relative effectiveness and validity of mood induction procedures: a meta-analysis," *European Journal of Social Psychology*, vol. 26, pp. 557–580, 1996.

[3] H. Gunes and B. W. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[4] A. M. Ruef and R. W. Levenson, "Continuous measurement of emotion: The affect rating dial," in *Handbook of emotion elicitation and assessment*, J. A. Coan and J. J. B. Allen, Eds. Oxford University Press, 2007, pp. 286–297.

[5] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELTRACE: An instrument for recording perceived emotion in real time," *ISCA Tutorial and Research Workshop on Speech and Emotion*, pp. 19–24, 2000.

[6] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–1, 2014.

[7] K. Diconne, G. K. Kountouriotis, A. E. Paltoglou, A. Parker, and T. J. Hostler, "Presenting KAPODI – The Searchable Database of Emotional Stimuli Sets," *Emotion Review*, vol. 14, no. 1, pp. 84–95, 2022.

[8] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[9] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[10] J. M. Girard, J. F. Cohn, L. Yin, and L.-P. Morency, "Reconsidering the Duchenne smile: Formalizing and testing hypotheses about eye constriction and positive emotion," *Affective Science*, vol. 2, no. 1, pp. 32–47, 2021.

[11] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 488–500.

[12] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proceedings of the 10th IEEE International Conference on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[13] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Apr. 2013, pp. 1–8.

[14] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, Jul. 2015.

[15] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and P. Maragos, "COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 54, 2017.

[16] R. Samide, R. A. Cooper, and M. Ritchey, "A database of news videos for investigating the dynamics of emotion and memory," *Behavior Research Methods*, vol. 52, no. 4, pp. 1469–1479, Aug. 2020.

[17] J. Rottenberg, R. D. Ray, and J. J. Gross, "Emotion elicitation using films," in *Hanbook of emotion elicitation and assessment*, J. A. Coan and J. J. B. Allen, Eds. Oxford University Press, 2007, pp. 9–28.

[18] J. M. Girard, "CARMA: Software for continuous affect rating and media annotation," *Journal of Open Research Software*, vol. 2, no. 1, p. e5, 2014. [Online]. Available: https://carma.jmgirard.com

[19] K. Mundnich, B. M. Booth, B. Girault, and S. Narayanan, "Generating labels for regression of subjective constructs using triplet embeddings," *Pattern Recognition Letters*, vol. 128, pp. 385–392, Dec. 2019.

[20] K. Kercher, "Assessing Subjective Well-Being in the Old-Old: The PANAS as a Measure of Orthogonal Dimensions of Positive and Negative Affect," *Research on Aging*, vol. 14, no. 2, pp. 131–168, 1992.

[21] A. Mackinnon, A. F. Jorm, H. Christensen, A. E. Korten, P. A. Jacomb, and B. Rodgers, "A short form of the Positive and Negative Affect Schedule: evaluation of factorial validity and invariance across demographic variables in a community sample," *Personality and Individual Differences*, vol. 27, no. 3, pp. 405–416, 1999.

[22] D. ten Hove, T. D. Jorgensen, and L. A. van der Ark, "Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs," *Psychological Methods*, Sep. 2022.

[23] J. M. Girard, "varde: An R package for variance decomposition," 2023. [Online]. Available: https://github.com/jmgirard/varde

[24] P. L. Smith and R. M. Luecht, "Correlated Effects in Generalizability Studies," *Applied Psychological Measurement*, vol. 16, no. 3, pp. 229–235, 1992.

[25] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[26] R. P. McDonald, *Test theory: A unified approach*. Erlbaum, 1999.

[27] D. B. Flora, "Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates," *Advances in Methods and Practices in Psychological Science*, vol. 3, no. 4, pp. 484–501, 2020.

[28] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012.

[29] T. D. Jorgensen, S. Pornprasertmanit, A. M. Schoemann, and Y. Rosseel, "semTools: Useful tools for structural equation modeling," 2022. [Online]. Available: https://CRAN.R-project.org/package=semTools

[30] D. McNeish, "Thanks coefficient alpha, we'll take it from here," *Psychological Methods*, vol. 23, no. 3, pp. 412–433, 2017.

[31] R. Rosenthal, "Conducting judgment studies: Some methodological issues," in *The new handbook of methods in nonverbal behavior research*, J. A. Harrigan, R. Rosenthal, and K. R. Scherer, Eds. Oxford University Press, 2005, pp. 199–234.

[32] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The Ambiguous World of Emotion Representation," Sep. 2019, arXiv:1909.00360 [cs].

[33] J. McAuley, *Personalized machine learning*. Cambridge University Press, 2022.

[34] D. Watson and A. Tellegen, "Toward a consensual model of mood," *Psychological Bulletin*, vol. 98, no. 2, pp. 219–235, 1985.

[35] J. M. Girard and A. G. C. Wright, "DARMA: Software for Dual Axis Rating and Media Annotation," *Behavior Research Methods*, vol. 50, no. 3, pp. 902–909, 2018.

[36] K. Fayn, S. Willemsen, R. Muralikrishnan, B. Castaño Manias, W. Menninghaus, and W. Schlotz, "Full throttle: Demonstrating the speed, accuracy, and validity of a new method for continuous two-dimensional self-report and annotation," *Behavior Research Methods*, vol. 54, no. 1, pp. 350–364, 2022.

[37] D. C. Rubin and J. M. Talarico, "A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words," *Memory*, vol. 17, no. 8, pp. 802–808, 2009.