Collecting Mementos

A Multimodal Dataset for Context-Sensitive Modeling of Affect and Memory Processing in Responses to Videos

Dudzik, Bernd; Hung, Hayley; Neerincx, Mark A.; Broekens, Joost

DOI
10.1109/TAFFC.2021.3089584

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Collecting *Mementos*: A Multimodal Dataset for Context-Sensitive Modeling of Affect and Memory Processing in Responses to Videos

Bernd Dudzik, *Student Member, IEEE*, Hayley Hung, *Member, IEEE*, Mark Neerincx, and Joost Broekens

Abstract—In this article we introduce *Mementos*: the first multimodal corpus for computational modeling of affect and memory processing in response to video content. It was collected online via crowdsourcing and captures 1995 individual responses collected from 297 unique viewers responding to 42 different segments of music videos. Apart from webcam recordings of their upper-body behavior (totaling 2012 minutes) and self-reports of their emotional experience, it contains detailed descriptions of the occurrence and content of 989 personal memories triggered by the video content. Finally, the dataset includes self-report measures related to individual differences in participants' background and situation (*Demographics*, *Personality*, and *Mood*), thereby facilitating the exploration of important contextual factors in research using the dataset. We describe 1) the construction and contents of the corpus itself, 2) analyse the *validity* of its content by investigating biases and consistency with existing research on affect and memory processing, 3) review previously published work that demonstrates the *usefulness* of the multimodal data in the corpus for research on automated detection and prediction tasks, and 4) provide suggestions for how the dataset can be used in future research on modeling *Video-Induced Emotions*, *Memory-Associated Affect*, and *Memory Evocation*.

Index Terms—Multimodal dataset, personal memory, video-induced emotion, memory evocation, memory-associated affect, affect detection, video affective content analysis, context-sensitivity, personalization

1 Introduction

Consuming video content is an essential part of peoples' everyday lives. It fulfills needs ranging from the merely practical – learning from recordings of educational material, such as tutorials or lectures –, towards the deeply socio-emotional [1] – watching home videos to commemorate a lost loved one, or forget about a stressful day by watching an entertaining movie with friends. Because of this broad relevance, research on Affective Computing actively explores approaches to automatically predict the emotional and cognitive effects that watching a given video produces in viewers. To make these predictions approaches typically either 1) analyze the audiovisual signals comprising a video's content [2], or 2) analyze sensor data describing viewers' behaviors and physiological processes. The resulting information about how people respond or process video content has potentially a great variety of applications. Examples include providing automatic feedback to content creators or enable applications involving media retrieval to respond to the needs of their users dynamically [3].

While existing research has primarily focused on predicting the immediate emotional impact of video viewing on individuals [4], efforts have also touched on the ebb and flow of viewers' attention while doing so [5], or the ability of content to be remembered [6]. Independent of the specific construct that is the target, publicly available datasets are an essential component for progress in research because they facilitate computational modeling and benchmarking [7].

In this paper, we introduce and describe *Mementos*: a novel dataset for modeling affect and memory processing occurring in viewers when they engage with video content. Concretely, it captures the feelings and personal memories triggered in a diverse audience while they are watching a series of music videos online. Additionally, it contains recordings of their behavior while doing so. We have used this corpus in previous research to model the contextual influence of occurring personal memories on the emotional impact of videos [8], [9], [10]. However, we believe that it can benefit future computational work on affect and memory processing more broadly, facilitating novel research beyond our initial inquiries. Motivated by this, we make the following contributions:

- *Presentation of a Multimodal Dataset:* We describe the design and contents of the first multimodal dataset that captures the occurrence and impact of viewers' personal memories on their emotional responses to video stimuli.

- *Bernd Dudzik, Hayley Hung, and Mark Neerincx are with the Department of Intelligent Systems (INSY), Delft University of Technology, 2628XE Delft, The Netherlands. E-mail: {b.j.w.dudzik, h.hung, m.a.neerincx}@tudelft.nl.*
- *Joost Broekens is with the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, 2333CA, Leiden, The Netherlands. E-mail: d.j.broekens@liacs.leidenuniv.nl.*

Manuscript received 5 Jan. 2021; revised 24 Apr. 2021; accepted 1 June 2021. Date of publication 16 June 2021; date of current version 31 May 2023. (Corresponding author: Bernd Dudzik.) Recommended for acceptance by M. CHETOUANI. Digital Object Identifier no. 10.1109/TAFFC.2021.3089584

- *Analysis of its Validity:* By presenting findings from a series of statistical analyses, we demonstrate that Mementos captures 1) a diverse and plausible set of affective responses, 2) effects and relationships that are consistent w.r.t. existing psychological research, and 3) multimodal data with sufficient quality for computational modeling.
- *Demonstration of its Usefulness:* We review and discuss the findings of two of our previous studies using Mementos for multimodal machine learning experiments to demonstrate the corpus' usefulness for this kind of research.
- *Suggestions for its Use in Future Research:* We provide suggestions for how Mementos may be useful for research on modeling *Video-Induced Emotions*, *Memory-Associated Affect*, and *Memory Evocation*.

Researchers can find instructions for requesting access to the dataset online: http://mementos-dataset.com/. Gaining authorization requires signing an End User License Agreement (EULA) to ensure compliance with the conditions under which participants provided their consent.

2 Motivation for Creating *Mementos*

Personal memories and past experiences are important drivers for emotional responses to situations, including interactions with media content. Empirical psychology has established both the capacity of media content to evoke personal memories readily [11], and the ability of such recollections to have a substantial emotional impact [12]. Moreover, findings suggest that the emotional impact of media stimuli on individuals matches their feelings towards the connected memories [13]. This ability to evoke emotional associations with our past is at the center of many media usage patterns (e.g., taking holiday photos or reminiscing over music from our teenage years). Relating to these memory-related uses is increasingly of interest to applications (see, e.g., [14]). Additionally, paying attention to triggered mental stimuli, such as thoughts about the past, is one source for individuals to lose engagement with tasks involving media [5]. As such, for technologies to intelligently support people in interactions with media content, they can benefit substantially from understanding when memories occur in viewers, what these memories are likely about, and how they will be emotionally experienced (see Dudzik *et al*. [15] for an in-depth discussion). Despite this, the evocative potential of stimuli and the emotional influence of personal memories have remained largely unexplored in computational research. Consequently, the primary motivation for the construction of Mementos is to 1) provide researchers with a corpus of multimodal data that captures the occurrence of personal memories in response to videos, 2) assesses their content, 3) and measures their impact on viewers' emotions. As far as we are aware, it is the first dataset on this topic.

In the following, we discuss a series of additional goals and constraints that influenced the design of Mementos, making it attractive for re-use in future research.

2.1 Responses Should be Ecological Valid

Represent Diversity of Viewers and Situations. Contemporary video content is consumed by a vastly diverse community of viewers, alone or in a group, and in a wide variety of circumstances [16]. Such differences in context are known to strongly affect both emotional experience and expression in general [17], particularly in response to media content [7]. Similar findings exist for the influence of context on the elicitation of personal memories [18]. Together, these findings indicate that how a particular viewer feels about a specific video (and whether memories play a role in it) may strongly depend on who they are and where they watch it. Moreover, similar feelings may manifest differently in terms of behavioral or physiological signals. For this reason, a dataset for modeling responses to video content must strive to adequately reflect the variation in viewers and situations under which such stimuli are encountered [7]. Awareness of the need for extensive and diverse corpora of responses to videos has motivated researchers to increasingly undertake data collection in an online setting (e.g., [19], [20], [21], and this is also the approach that we use for the construction of Mementos. In particular, all the self-reports and behavioral recordings it contains are collected using a web-based procedure that imposes only a minimal set of restrictions on who can participate and the circumstances in which they can do so. Consequently, Mementos is likely to possess an overall high degree of ecological validity regarding these aspects.

Include Emotionally Ambiguous Video Stimuli. Traditionally, video material for emotion induction is selected to elicit pronounced and homogeneous responses across viewers, both for experiments in psychology (e.g., *EMDB* [22]), as well as in databases for affect modeling in computer science (e.g., *DEAP* [23], and *AMIGOS* [24]). However, filtering out material eliciting ambiguous responses results in a set of stimuli that is not representative of content that viewers engage with throughout their everyday lives. In particular, responses to these examples are abnormally content-driven (e.g., by spanning extreme topics) and thus suppress the substantial influence that situation- and person-specific effects can have on the subjective emotional experience of video content [7]). Not capturing such influences in a dataset for predictive modeling is a serious limitation on its ability to facilitate the development of reliable technology because findings derived from it may not generalize beyond its set of artificial examples. For this reason, an additional motivation for creating Mementos is to provide a dataset that explicitly selects a set of videos that is balanced for its ability to elicit both pronounced and ambiguous responses from participants (see Section 3.2).

Use In-the-Wild Recording Conditions. Apart from limiting variation in terms of context, collecting datasets in a laboratory typically has the additional effect of fixing the technical quality of audiovisual recordings. In particular, creators typically optimize for future analysis (e.g., by controlling lighting conditions and removing occlusions). However, these recording conditions are unrealistic for data available to applications deployed *In-the-Wild* and can lead to an unexpected and poor performance of machine analysis. Because recordings of viewers' behavior in Mementos are collected from their webcams and with minimal restrictions on environmental conditions, they are highly representative of the technical conditions that automatic analysis would face in many real-world applications.

TABLE 1
Comparison of Databases for Video-Induced Affect

Database		Protocol		Stimuli		Response		Context Measures		
Name	Type	N_P	Setting	N_S	Content	Beh.	Phys.	Dem.	Pers.	Additional
AMIGOS [25]	VC	40	Lab	20	Films	✓	✓	✓	✓	Mood, Social Presence
ASCERTAIN [26]	VC	58	Lab	36	Films	✓	✓	✓	✓	
CP-QAE-I [27]	VC	76	Online	12	Films			✓	✓	Video Quality
DEAP [24]	VC	32	Online	120	Music Vids	✓†	✓†	✓†		
DECAF [28]	VC	30	Lab	36	Films	✓	✓	✓		
LIRIS-ACCEDE [5]	SC	N/A*	Online	9800	Films					
MAHNOB-HCI [29]	VC	27	Lab	20	Films	✓	✓	✓		
VIDEO EMOTION [30]	SC	N/A*	Lab	1101	Soc. Media					
SEWA [31]	VC	398	Lab	4	Adverts	✓		✓		
Mementos	VC	297	Online	42	Music Vids	✓		✓	✓	Mood, Memories

VC: *Viewer-Centric Corpus;* SC: *Stimulus-Centric Corpus;* N_P*: Number of Participants;* N_S*: Number of Stimuli;* Beh.*: Data on Behavior;* Phys.*: Data on Physiological;* Dem.*: Data on Participants' Demographics;* Pers.*: Data on Participants' Personality*

*Not applicable, since these corpora focus on video-level aggregates
† Data was collected only for a subset of participants in a Lab.

2.2 Relevant Context Variables Should be Measured

In addition to measuring responses across different contexts (i.e., ecological validity), it is also desirable that corpora for affect modeling provide detailed data about these variations [31]. Not only do these measures provide insights into potential limitations and harmful biases that a dataset may suffer from, but it is also information that can be essential for research on personalized or context-sensitive approaches for predicting responses to videos. To address this aspect, Mementos contains information about viewers that has been identified as accounting for individual differences in affective responses: 1) *Demographics*, 2) *Personality*, and 3) *Mood*.

Demographic information is often capable of capturing broad similarities and differences in people's past experiences, attitudes, and behaviors. Notably, findings show that the intensity of viewers' emotional experience to video stimuli differs depending on their age [32]. Similarly, people may respond differently to video content, depending on the cultural values of the country that they are nationals of [33]. Moreover, personality traits provide broad insights into individual differences between people and explain variation in affective responses to videos [33]. In contrast to emotions, moods are enduring, low-intensity affective states that are typically not directed towards a specific event or stimulus [34]. Nevertheless, they can exercise a broad influence on individuals' experience and behavior in a given situation, including affective responses to videos [7].

2.3 Creation Should Support Interdisciplinary Work

Affective Computing involves computational modeling of cognitive, affective, and social processes, often focusing on supporting human-computer interactions. As such, it is a technological enterprise that not only heavily relies on domain knowledge from psychology and the social sciences but that also has the potential to make substantial contributions to research in these fields [35]. Such contributions can include the collection and sharing of corpora for analysis and modeling. However, two important challenges hamper such interdisciplinary exchanges: 1) different goals in data collection processes [36], and 2) the accurate representation of psychological and social constructs [35]. In particular, corpora in computer science are typically collected with a strong focus on rich and technologically valid sensor data for automatic processing and analysis but sometimes model psychological constructs in an ad-hoc fashion. In contrast, researchers in the social sciences or psychology create text, speech, or video corpora often with manual extraction of information in mind, and the focus of their design rests heavily on validity and experimental control.

To foster interdisciplinary use, we designed Mementos to balance technologically sound data for automatic analysis with capturing psychological constructs in a psychologically grounded fashion. For example, we measure individuals' affective responses in terms of the widely used *Pleasure-Arousal-Dominance (PAD)* framework [37], using the Affect-Button, a well-validated measurement instrument [38]. It quantifies affective states and judgments in terms of the three dimensions of *pleasure (P)* (is an experience pleasant or discomforting?), *arousal (A)* (does it involve a high or low degree of bodily excitement?), and *dominance (D)* (does it involve the experience of high or low control over the situation?). This representation is ideal for fostering cross-disciplinary use since it is prominent in Affective Computing research and psychology (e.g., IAPS [39]).

2.4 Related Work

2.4.1 Databases of Video-Induced Affect

A range of datasets for modeling affective responses to videos is publicly available to the research community. Here we review relevant examples to highlight the unique contributions of Mementos (see Table 1 for an overview). For this purpose, we differentiate between corpora that are either 1) *Stimulus-Centric (SC)* or 2) *Viewer-Centric (VC)*, depending on the motivation for their creation. The former type focuses on collecting affective self-reports about many different examples of video content, but from comparatively few viewers for each and often with no additional information about their behavior or context. These corpora are typically

geared towards Video Affective Content Analysis [2], i.e., analysis of the audiovisual content of a video to automatically predict the emotions it is expected to induce in viewers [4]. Additionally, affect is often labeled at the video level, e.g., through aggregating ratings for the same stimulus. Principal examples include *LIRIS-ACCEDE* [19], or *VideoEmotion* [29]. In contrast, corpora focusing on viewers rely on a comparatively small set of videos for emotion induction to capture self-reports and multimodal measures in response to each from a larger pool of individuals. They are primarily used for work on *Multimodal Affect Detection* [40], i.e., analyzing behavioral, physiological, and sometimes contextual data to predict the emotional response of individuals. Relevant examples include *DEAP*[23], *DECAF*[27], *MAHNOB-HCI*[28], *AMIGOS* [24], *ASCERTAIN* [25], and *SEWA* [30]. Noteworthy is also *CP-QAE-I* [26], which does not contain behavioral measures, but provides rich context about individual viewers' background. Importantly, VC databases can, in principle, serve to model cross-video differences (i.e., through the video-wise aggregation of responses). For example, DEAP, DECAF, or MAHNOB-HCI have been designed explicitly with this perspective in mind. However, in practice, these corpora are less suited to do so than specialized SC corpora because of their comparatively small amount of stimuli.

According to the above categorization, Mementos can be considered as a viewer-centric dataset. It contains responses to a comparable amount of videos to AMIGOS, ASCERTAIN, DEAP, DECAF, and MAHNOB-HCI, but from a much larger participant pool. Like these corpora, Mementos provides recordings of viewers' behaviors. However, unlike them, it does not offer physiological measures for analysis. It was not collected under laboratory conditions, where it is more feasible to take such physiological measures. VC corpora typically collect at least demographic information to contextualize participants' responses, with an increasing number also accounting for personality. However, only AMIGOS and CP-QAE-I offer a comparable range of relevant contextual factors. In contrast to SEWA and CP-QAE-I, which are specifically constructed for cross-cultural comparisons, this was not a primary goal underlying Mementos. Finally, Mementos is the only corpus capturing memories triggered by the video stimuli used for emotion elicitation.

2.4.2 Databases of Memory Processing

Human Memory Processing can be broadly divided into three distinct components: memory encoding (what is stored?), retention (what is forgotten?), and retrieval (what is accessed?). Moreover, retrieval can be initiated in different ways, either voluntary (i.e., we intentionally remember something) or involuntary (i.e., we are spontaneously reminded of something by an internal or external cue). The memory processing targeted by Mementos is retrieval that is involuntarily initiated by videos in participants exposed to them. To the best of our knowledge, it is the only publicly available dataset for multimodal modeling of involuntary retrieval collected in the wild.

However, a few corpora exist that support computational research on memory processes related to video material. One type focuses on viewers' encoding of video content, i.e., its *memorability*. Here participants are first exposed to some video content and then asked to report what they remember of it at a later point in time. Noteworthy examples include the corpus developed by Samide *et al.* [41] and the dataset used for the Memorability-task at MediaEval. [42]. In addition, there is computational research that is closely related to modeling involuntary memory retrieval, studying attentional shifts between external stimuli (e.g., video content) and internal stimuli (e.g., thoughts or memories) during media consumption [5]. However, data collection for this paradigm is difficult and requires careful experimental settings, and – to the best of our knowledge –, no publicly available corpora exist.

3 Data Collection Framework

In this section, we provide a detailed description of the design and execution of the online study through which we collected the data forming the contents of Mementos.

3.1 Participant Selection

We limited participation to individuals capable of understanding and speaking English. Further, we request that they undertake the entire online study in a calm environment and give their undivided attention to it. Moreover, participants need to use a laptop or desktop computer (i.e., no mobile or tablet) with a functioning webcam and participate in lighting conditions in which their face remains visible. Similarly, they have to ensure that they are the only person in the recordings, i.e., no other individuals visible in the background. Finally, we restricted their age to the range of 25 to 46 years. We enforce this constraint to align the age of music videos selected for evoking responses in our study (see below) with years that fall into a period in participants' life between the age of 15 to 30. This age range is associated with exceptionally accessible personal memories, a phenomenon labeled in psychological theory as the *reminiscence bump* [43]. The idea behind this alignment is to maximize the capacity of our stimuli to trigger personal memories in viewers.

3.2 Measures and Materials

3.2.1 Video Stimuli for Evoking Responses

For evoking affective and memory responses, we rely on a subset of the music video stimuli part of the *DEAP dataset* [23]. Each segment has a length of 60 seconds and is extracted from the overall clips. We decided to select from this corpus for two reasons: First, because existing findings highlight the potency of music for triggering emotional memories in listeners (see, e.g., the findings of Janata *et al.* [12]). Second, the corpus contains ratings for the emotional impact stimuli in terms of the PAD framework from multiple viewers. These ratings provide us with insights into the expected distribution of emotional responses to the videos, which we use for balancing purposes when selecting for our study.

From the 120 video segments comprising the *Online subjective annotation*-part of the DEAP corpus, we select 42 videos for evoking responses. We choose stimuli based on their variation for the pleasure, arousal, and dominance they

TABLE 2
DEAP Video Stimuli Selected for Emotion Elicitation

ID	Source	Genre*	Release
1	Alphabeat - Fascination	Pop	2007
2	Emilíana Torrini - Jungle Drum	Pop	2008
3	The Go! Team - Huddle Formation	Pop	2004
8	4 Strings - Let It Rain (Dj 4 Strings Vocal Mix)	Electronic	2003
12	Mika - Love Today	Pop	2007
13	I'm From Barcelona - We're From Barcelona	Pop	2006
17	Grand Archives - Miniature Birds	Pop	2008
18	Jack Johnson - Breakdown	Pop	2005
23	Oren Lavie - Her Morning Elegance	Pop	2007
24	Bright Eyes - First Day Of My Life	Rock	2005
28	Lara Fabian - Tango	Pop	2001
32	Gary Jules - Mad World	Pop	2001
33	Wilco - How To Fight Loneliness	Rock	1999
37	The Submarines - Darkest Things	Rock	2006
41	James Blunt - Goodbye My Lover	Pop	2004
44	A Fine Frenzy - Goodbye My Almost Lover	Pop	2005
45	Kings Of Convenience - The Weight Of My Words	Rock	2001
48	Limp Bizkit - Break Stuff	Rock	1999
49	Parkway Drive - Smoke 'Em If Ya Got 'Em	Rock	2005
54	Blitzkid - Nosferatu	Rock	2006
55	Grace Jones - Corporate Cannibal	R&B	2008
56	Dead To Fall - Bastard Set Of Dreams	Rock	2004
57	Trapped Under Ice - Believe	Rock	2009
59	Stigmata - В отражении глаз	Rock	2009
63	Blur - Song 2	Rock	1997
66	Beastie Boys - Sabotage	Rap	1994
70	Blink 182 - First Date	Rock	2001
71	Europe - The Final Countdown	Rock	1986
72	Benny Benassi - Satisfaction	Electronic	2003
81	Black Eyed Peas - My Humps	Rap	2005
83	Manu Chao - Me Gustas Tu	Pop	2001
85	Taylor Swift - Love Story	Pop	2008
86	Pink Floyd - Marooned	Rock	1994
90	Nouvelle Vague - Dancing With Myself	Pop	2006
91	Moby - Why Does My Heart Feel So Bad	Electronic	1999
99	Requiem For A Dream - Ending Scene	Classical	2000
101	Portishead - Roads	Pop	1994
111	Napalm Death - Procrastination On The Empty Vessel	Rock	2009
112	Sepultura - Refuse Resist	Rock	1993
114	Deicide - Homage For Satan	Rock	2006
116	Dark Funeral - My Funeral	Rock	2009
120	Arch Enemy - My Apocalypse	Rock	2005

ID: *Number assigned in the DEAP dataset*.

**Based on AllMusic.com*

evoke in viewers. Concretely, we try to balance more emotionally ambiguous stimuli with less ambiguous ones by selecting an equal amount of videos per affective dimension that possess either a high- or a low- degree of variation. See Table 2 for a description of the selected video stimuli, including the title, release year, and genre.

3.2.2 Self-Report Measures

Viewer-Specific Measures. We collect the following self-reports to capture relevant aspects of participants' backgrounds, i.e., they are obtained once per viewer.

- *Demographics:* We capture self-reports of participants' age in years, their gender, and nationality.
- *Personality:* We measure viewers' personality in terms of the HEXACO scheme, which comprises six orthogonal trait-dimensions: *Honesty-Humility (H)*, *Emotionality (E)*, *eXtraversion (X)*, *Agreeableness (A)*, *Conscientiousness (C)*, and *Openness to experience (O)*. For assessing viewers we rely on the *Brief HEXACO Inventory (BHI)* [44], which has been designed for a quick assessment (it consists of only 24-items) while minimizing the loss to validity. This makes it particularly suitable for deployment in crowd-sourcing scenarios.
- *Mood:* We quantify mood in terms of pleasure-, arousal- and dominance-ratings on a continuous scale constrained to the interval of $[-1, +1]$. We obtain ratings with the *AffectButton* [38] instrument – an interactive widget displaying an iconic facial expression that changes in response to mouse or touch interaction. It enables users to select the facial expression that matches their affective judgment most closely. The benefits of this instrument are 1) that it facilitates PAD-ratings without prior knowledge of the dimensions and the underlying psychological framework, and 2) that it requires minimal time for providing them. For data collection in Mementos, the AffectButton Widget had a size of $240 * 240$ pixel. With these settings, the instrument facilitates 220 unique inputs along the X and Y-axes in a $[-1, 1]$ interval each (see Broekens & Brinkman [38] for a detailed description of the mapping to PAD ratings and a validation study).

Response-Specific Measures. We collect the following self-reports to describe participants' responses to a specific video stimulus, i.e., they are taken once for a specific viewer's response to a particular video.

- *Induced Emotions:* We capture viewers' ratings for their emotional response to a video with the AffectButton.
- *Familiarity:* We ask participants to describe the degree to which they had previously been exposed to a video. We hypothesized that familiarity influences the chance of videos to trigger associated memories in individuals. Ratings use a 5-point Likert-Scale in the interval $[0, 4]$, matching the labels: {*"Never"*, *"Once"*, *"A few times"*, *"Often"*, *"Very Often"*}.

Memory-Specific Measures. In the following, we describe measures that we deploy to capture relevant qualities of any personal memories that viewers recollect.

- *Memory Content:* To capture the content of personal memories, we ask participants to *(1)* describe these in a short free-text (*Memory Description*), and *(2)* rate their age in the memory from a list of predefined ranges: {*"1-10 years"*, *"11-20 years"*, *"21-30 years"*, *"31-40 years"*, *"41-50 years"*}.
- *Memory-Associated Affect:* We measure how people feel about the content of the personal memories that videos trigger in them. They provide ratings in terms of pleasure, arousal, and dominance using the AffectButton instrument. Moreover, participants label their feelings with up to three free-text labels of their choice.

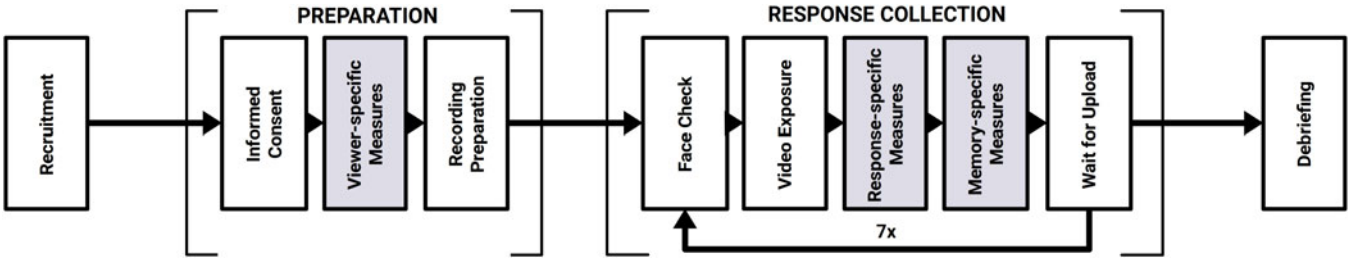

Fig. 1. Protocol for data collection in our online study from participating crowd-workers. Purple fields refer to stages at which we collected the respective Self-report Measures listed in Section 3.2.

- *Memory Experience:* We collect information about two qualitative aspects of participants' recollective experience that may influence memories' emotional impact: *(1)* the clarity and intensity with which they experience the memory (*Vividness*), and 2 how connected it is to the video that has triggered it (*Connectedness*). For assessment, we deploy a custom slider-based rating instrument. Moving the bar of the widget results in ratings bound to the interval $[1, 100]$. For vividness, we labeled the extremes of this scale *"Not vivid at all"* and *"Very Vivid"*, while for connectedness they are *"Not connected at all"* and *"Very Strongly Connected"*.

3.2.3 Webcam Recordings

We capture visual recordings of participants' faces at 30 frames per second and a *minimum* resolution of $640 * 480$, and audio input with a sampling rate of 44100 Hz.

3.2.4 Online Application for Data Collection

For collecting data from participants we developed a specific online application based around the JavaScript-framework *jsPsych*[1] [45], which they can access through their browser.

It guides them through the entire online study, presenting them with a random selection of our selected video stimuli and the survey elements necessary for the self-report measures (see Section 3.3 for details about the protocol). Additionally, it handles the recording and storage of face recordings with participants' webcams. We implemented a mechanism giving priority to videos with the least amount of responses so far when selecting a sample for participants. This mechanism helps collect a roughly equal amount of responses for each video, even in cases where crowd-workers fail to complete the entire protocol (e.g., due to technical problems). Furthermore, the application is capable of automatically detecting the presence of participants' faces in their webcam feed using the JavaScript-based face tracker *pico.js*.[2] These detections are solely used by the application to provide participants with feedback for creating suitable recording conditions. Finally, we implemented several mechanisms to ensure that individuals pay their undivided attention to the study. For example, we present them with warning messages if they navigate away from the browser window in which the application is running.

1. https://www.jspsych.org
2. https://github.com/nenadmarkus/picojs

3.3 Protocol

In the following, we describe the different phases of the protocol of our online study for data collection. See Fig. 1 for a graphical overview.

3.3.1 Recruitment-Phase

The study was announced to crowd-workers on the Amazon Mechanical Turk platform. Interested crowd-workers could participate in the study by following a link to the online application. In total, we recruited 300 individuals, compensating each one for their participation with a sum of 6 *USD* upon completion.

3.3.2 Preparation-Phase

Acquisition Informed Consent. All participants are required to provide their informed consent before entering the actual study, both regarding the tasks involved and the usage of their data.

Acquisition Viewer-Specific Measures. Next, participating crowd-workers fill out a survey containing the viewer-specific measures. On separate slides, the application first requests information about their basic demographics, then provides them with the personality survey, and finally requests a rating of their mood.

Recording Setup. At this stage, the web application guides participants through the process of setting up acceptable recording conditions. To this end, it presents them with the input of their webcam and suggestions for ensuring good quality. Participants can only continue if the application's face tracking algorithm can successfully detect their faces. Then participants are presented with a test video together with the instructions for a correct audio setup.

3.3.3 Response Collection-Phase

With preparations concluded, the application chooses a sample of 7 videos from our pool for presentation to the participant. Then the following steps are repeated once for each video in this selection.

Face Check. In this phase, we use the application's face tracker to ensure that participants face in the image, preventing continuation if it is not. We provide participants with feedback about the success of the tracking and a preview of the video stream to adjust their recording conditions (e.g., lighting).

Video Exposure. We present a random video from the sample drawn for them at the beginning of the response collection phase to the participant. Playback starts automatically and does not allow for pausing or rewinding.

Acquisition Response-Specific Measures. After playback has concluded, participants report how the video made them feel and their previous exposure to the stimulus.

Acquisition Memory-Specific Measures. Next, we instruct participants to reflect on their viewing experience and report any memories that they had recollected. Because it is plausible that a video triggers multiple memories throughout its duration, we set no upper limit to how many they can report. However, we remind them only to report memories *(1)* if they have experienced them, and *(2)* have done so during exposure to the video. Independently of whether they report memories or not, all participants have to spend a minimum of 90 seconds in this stage before they can continue. This measure aims to discourage crowd-workers from minimizing the time spent on their participation in the study by not reporting memories that they have recollected.

Waiting for Upload. Depending on participants' internet connection, uploading their webcam recording may take longer than capturing the self-report measures to a video. In this case, they have to wait before seeing the next video.

3.3.4 Debriefing-Phase

After completing the response collection phase, the application informs participants of their successful completion of the study. It provides them with a unique code to claim their compensation through the Mechanical Turk platform and contact information for further requests.

3.4 Ethics Statement

The procedures for collecting and sharing the dataset were approved by the university's Human Research Ethics Committee (*ID: 658*).

4 Dataset Curation and Contents

Through our online data collection, we managed to acquire a *Raw Dataset* consisting of a total of $N = 2098$ individual responses from $N = 300$ participating crowd-workers. In this section we outline *(1)* how we processed this data to create the definitive version that we are publishing for use by the research community (*Curated Dataset*, $N = 1995$ responses from $N = 297$ unique viewers), and *(2)* descriptive statistics of its contents.[3]

4.1 Curation

4.1.1 Data Cleaning and Processing

Self-Report Measures. As part of creating the curated version of the dataset for release to the research community, we applied the following operations to the collected self-reports:

- *Computing PAD-Intensity Scores:* We added a single metric for the intensity of each of the PAD ratings in our dataset (i.e., Mood, Induced Emotion, and Memory-associated Affect). Inspired by findings from Reisenzein [46], we represent intensity as the magnitude of ratings in terms of PAD-scores, using the following formula:

$$I = \frac{1}{\sqrt{3}} \sqrt{\left(p^2 + \left((a+1)/2\right)^2 + d^2\right)}. \quad (1)$$

  Here p, a, and d are the pleasure, arousal and dominant components of a particular rating. Importantly, we interpreted negative arousal values as low intensity, motivated by the layout of the AffectButton instrument, which maps maximum negative arousal to neutral face representations in the centre of the widget [38].
- *Extract Text Complexity:* We calculate two measures to characterize the complexity of free-text memory descriptions: the first is a *Word Count (WC)*, denoting the total number of words in a description. The second is the *Flesch Reading Ease score (FRES)*. It is a widely used metric to quantify the readability of texts using their average sentence length and average number of syllables in its calculation [47]. High scores denote simple sentences that are easy to read (with a maximum of 121), while low scores demarcate complex sentences that are hard to read (arbitrary minimum).

Webcam Recordings. Similarly, we applied the following processing and feature extraction steps to the raw behavioral recordings to create the curated dataset.

- *Transcoding Webcam Recordings:* The vast majority of the raw footage collected from participants was submitted with the minimum required resolution of $640 * 480$ $(2082/2098)$, with only a few instances of recordings in $1280 * 720$ $(16/2098)$. For a standardized analysis dataset, we transcode all raw footage to the majority resolution of $640 * 480$ and a frame rate of 30 frames per second.
- *Extracting Descriptors for Lighting Conditions:* We extract frames at a rate of 1 Hz from the webcam recordings in the dataset and convert them to grayscale images. To represent a recording's *Brightness*, we first average the pixel intensities within each of its frames and then average this across all of them. Similarly, we quantify *Contrast* by calculating the standard deviation of the pixel intensities in each recording's frames and then take the average across this.
- *Extracting Descriptors for Facial Expressions:* To capture information about the facial expressions of participants in the webcam recordings, we deployed the software *OpenFace 2.0* [48]. It provides an automatic coding of facial configurations according to a subset of the *Facial Action Coding System (FACS)*. This scheme decomposes activation of the combination of 45 individual muscles as distinct *Action Units (AUs)*. Concretely, OpenFace provides distinct intensity values for the activation of 17 AUs per frame (each value in the range $[0-5]$, where 0 denotes no activation). For description and analysis in this article, we summarize the coding extracted for each frame in a

3. Because of the repeated-measures design of our protocol, responses are not independent. To account for this, all statistical tests that we present in this article (e.g., ANOVAs and t-tests) use *Linear Mixed-Effects models (LMEs)* that include participants' identity as a random-intercept. We explicitly specify analyses for which this is not the case.

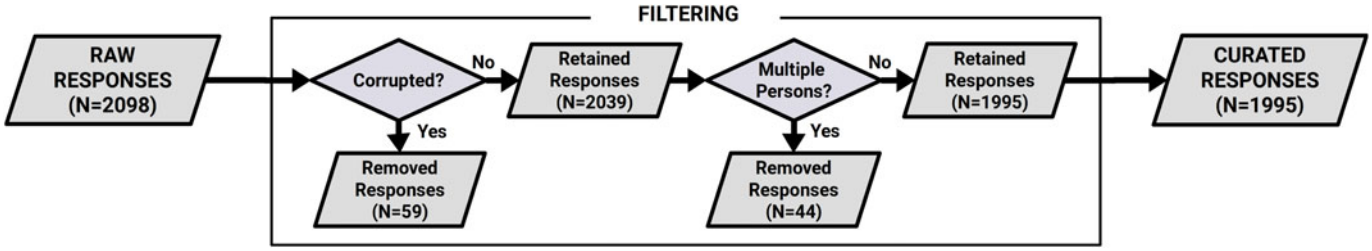

Fig. 2. Filtering of invalid responses from the dataset.

given recording by calculating two additional measures. The first is the *Average Maximal Action Unit Intensity*, for which we compute the maximum overall intensity values for each frame in a recording and aggregate them by taking the mean. As a second measure, we compute the *Average Presence of Facial Action* in a recording by taking the maximum over the intensity values for all AUs per frame and then calculate the proportion of frames for which this value is equal or exceeding 1 in the recording.

4.1.2 Filtering

Following the above preprocessing steps, we removed any responses from the dataset where either a component of the self-report measures or the webcam recordings were invalid, resulting in an incomplete record. A graphical overview of the sequential steps in this filtering process and the number of records removed by them is present in Fig. 2.

Removing Corrupted Recordings. Some responses from participants include webcam recordings that are technically corrupted in different ways, rendering them unsuitable for processing or analysis. The most common form of this includes substantial differences in recording duration from the expected 60 seconds matching our music videos. One potential reason for this is that slow connections of some participants result in longer exposure phases. We filtered out any responses with recordings outside of a range between 50 to 70 seconds for the curated dataset. Moreover, several recordings were not readable or contained only black frames and were also removed at this stage. In total, this resulted in the removal of 59 responses, leaving a total of 2039 remaining for further processing.

Removing Multiple Person-Recordings. Initial visual inspection of the recorded webcam material identified cases in non-participants are visible in the background. For example, in some cases, crowd-workers undertook the experiment in a public setting (e.g., an internet cafe) or shared their screen with other viewers. To enforce the constraint for isolated viewing across responses and systematically safeguard these bystanders' privacy, we attempt to filter out any responses with such multi-person recordings. For this purpose, we use the software *OpenPose*[4] [49] to automatically detect frames in the webcam recordings in which multiple people are visible. For any recording in which we detect at least one such frame, we undertake a manual inspection at 5-second intervals. We remove any video for which this reveals a visible person in the background. To preserve the ecological validity for technological challenges, we keep recordings that are suspect because a TV is running in the background or where photographs and posters with people in them are visible. This filtering removed a total of $N = 44$ from the remaining responses, resulting in a total of $N = 1995$ responses retained in the curated form of the dataset.

4. https://github.com/CMU-Perceptual-Computing-Lab/openpose

4.2 Statistics for Collected Self-Report Data

This section provides a descriptive overview and discussion of the collected self-report data contained in the dataset after processing and filtering (see Table 3 for summary statistics).

4.2.1 Viewer-Specific Measures

Demographics. The greatest part of the 297 remaining participants in the curated dataset reported being nationals of the United States of America ($N = 240$), followed by a substantial group from The Republic of India ($N = 45$). The small group of remaining participants ($N = 12$) hailed from a variety of different countries. Our sample covers the full range of ages that we targeted (24 to 46 years) but is leaning towards younger people ($M(SD) = 33.06(6.00)$). While our sample overall is relatively balanced w.r.t. gender ($N_{female} = 138$, $N_{male} = 150$), there is a greater imbalance for participants from India ($N_{female} = 11, N_{male} = 35$).

Personality. Except for Emotionality, our sample covers the entire range of possible scores for each HEXACO-trait (i.e., [0,4]). A one-way ANOVA with linear models reveals that scores differ significantly across the traits ($F(5, 1776) = 66.50$, $p < .001$). While the mean of scores for Emotionality and Agreeableness is located around the middle of the scale, scores for the remaining dimensions are substantially different from it (H: $t(296) = 14.93$, $p < .001$; X:$t(296) = 11.64$, $p < .001$; C: $t(296) = 15.38$, $p < .001$; O: $t(296) = 15.38$, $p < .001$). This systematic bias in personality scores indicates that we recruited participants leaning towards being socially confident, goal-oriented, and open to new aesthetic experiences.

Mood. Overall, participants, undertook the study in mood states leaning towards the positive, both in terms of experienced pleasure ($M(SD) = 0.45(0.4)$) and dominance ($M(SD) = 0.38(0.47)$). The distribution of arousal for mood ratings is strongly bi-modal, displaying distinct peaks for both arousal scores with positive polarity ($M(SD) = 0.69(0.30)$) and negative polarity ($M(SD) = -0.74(0.29)$). This is a known effect of the AffectButton rating instrument (see Broekens and Brinkman [38] for a discussion).

4.2.2 Response-Specific Measures

Induced Emotion. Similarly to the mood scores of participants, their emotional responses to the videos tend to be

TABLE 3
Statistics of Self-Report Data for Responses to Videos (Processed and Filtered Dataset)

Variable	Measure	$M\ (SD)$	Min/Max
Personality $N = 297^{+}$	**Hon/Hum.**	2.67 (0.75)	0.5/4.0
	Emotional.	1.94 (0.77)	0.0/3.75
	Extravers.	2.52 (0.76)	0.0/4.0
	Agreeabl.	2.06 (0.66)	0.25/4.0
	Conscien.	2.63 (0.7)	0.5/4.0
	Openness	2.76 (0.67)	0.0/4.0
Mood $N = 297^{+}$	**Pleasure**	0.42 (0.4)	-0.75/1.0
	Arousal	-0.12 (0.77)	-1.0/1.0
	Dominance	0.38 (0.47)	-1.0/1.0
	Intensity		
Demographics $N = 297^{+}$	**Age**	33.06 (6.01)	25.0/46.0
		$Unique$	$Top\ (Freq)$
	Nationality	3	USA (240)
	Gender	2	male (159)
Ind. Emotion $N = 1995^{*}$	**Pleasure**	0.2 (0.53)	-1.0/1.0
	Arousal	-0.12 (0.79)	-1.0/1.0
	Dominance	0.15 (0.58)	-1.0/1.0
	Intensity	0.52 (0.27)	0.01/1.0
Familiarity $N = 1995^{*}$	**Prev. Expo.**	0.26 (0.16)	0.2/1.0
Mem. Content $N = 989^{\dagger}$	**Descr. (WC)**	22.61 (13.38)	2/89
	Descr. (FRES)	78.0 (16.77)	-8.73/119.19
		$Unique$	$Top\ (Freq)$
	Age in Mem.	5	11-20y (437)
Mem. Affect $N = 989^{\dagger}$	**Pleasure**	0.33 (0.53)	-1.0/1.0
	Arousal	0.01 (0.78)	-1.0/1.0
	Dominance	0.29 (0.57)	-1.0/1.0
	Intensity	0.58 (0.26)	0.03/1.0
Mem. Exp. $N = 989^{\dagger}$	**Vividn.**	0.64 (0.27)	0.01/1.0
	Connect.	0.55 (0.32)	0.01/1.0

**Response-specific: measured once per response to a video*
**Viewers-specific: measured once per viewer*
†Memory-specific: measured once per memory

M (SD): Mean and Standard Deviation;
Min/Max : Range of values occurring;
Unique: No. of distinct categories;
Top(Freq): Category with the most items and their frequency count.

pleasurable ($M(SD) = 0.2(0.53)$) and score positive for dominance ($M(SD) = 0.15(0.58)$). Additionally, the distribution of self-reported induced arousal is bi-modal with clear peaks for values in the positive ($M(SD) = 0.70(0.29)$) and the negative range ($M(SD) = -0.77(0.29)$).

Familiarity. Overall viewers are largely unfamiliar with the video content that we have selected ($M(SD) = .26(0.16)$).

4.2.3 Memory-Specific Measures

In total, we collected 989 memories from 257 unique participants. During nearly half of all responses, viewers experienced recollections with at least one personal memory ($N = 935$). While participants had the option to report as many memories as they had experienced, only about 6 percent of all recollections ($N = 52$) involved more than 2 of them.

Memory Content. In Fig. 4 we provide an impression of the detail of participants' memory description in terms of their word counts and FRES, together with examples. Overall, descriptions of their memories are fairly long (word count: $M(SD) = 22.83(13.55)$), and use comparatively simple language (FRES: $M(SD) = 78.0(16.77)$, approx. readable by a pupil in 7th grade). Moreover, reported memories cover events throughout participants' lifespans, with a majority ($N = 437$) from a time when they were between 11 to 20 years old.

Memory-Associated Affect. On average, memories evoked in participants are pleasurable ($M(SD) = 0.33(0.53)$) and positive in dominance ($M(SD) = 0.29(0.57)$). Ratings for arousal in memory-associated affect are more diverse, also displaying a bi-modal pattern (positive peak: $M(SD) = 0.73(0.27)$; negative peak: $M(SD) = -0.74(0.29)$).

Memory Experience. While displaying a diversity, participants' recollective experience leaned more towards vivid than non-vivid recollection ($M(SD) = 0.64(0.27)$. The memories that videos evoked in participants were often not experienced as directly connected to the video that triggered them ($M(SD) = 0.55(0.32)$.

4.3 Statistics Recorded Behavior

Here we provide a brief overview and discussion of the behavioral recordings captured from participants (see Table 4 for summary statistics).

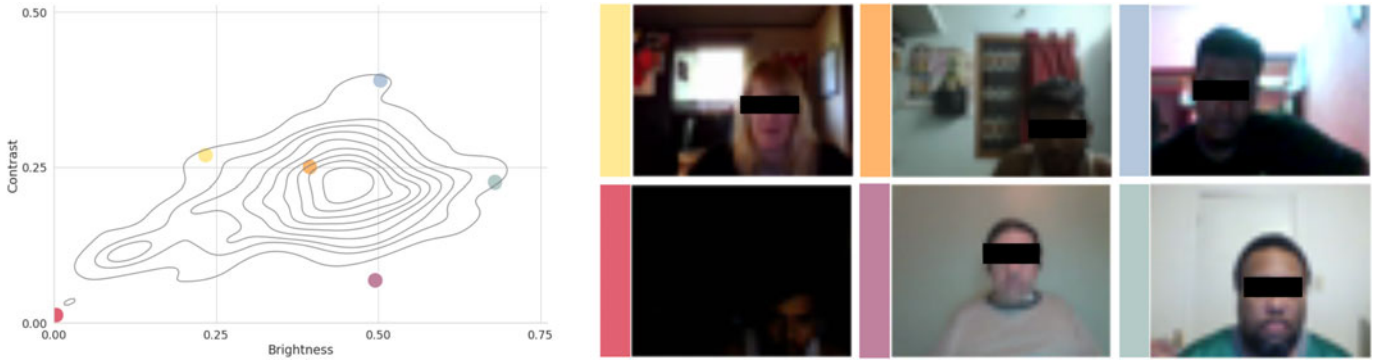

Fig. 3. Contour plot of the average brightness and contrast of the frames in recordings collected from participants. Images on the right are example frames taken from the recordings at the marked locations (down-sampled and masked to preserve participants privacy).

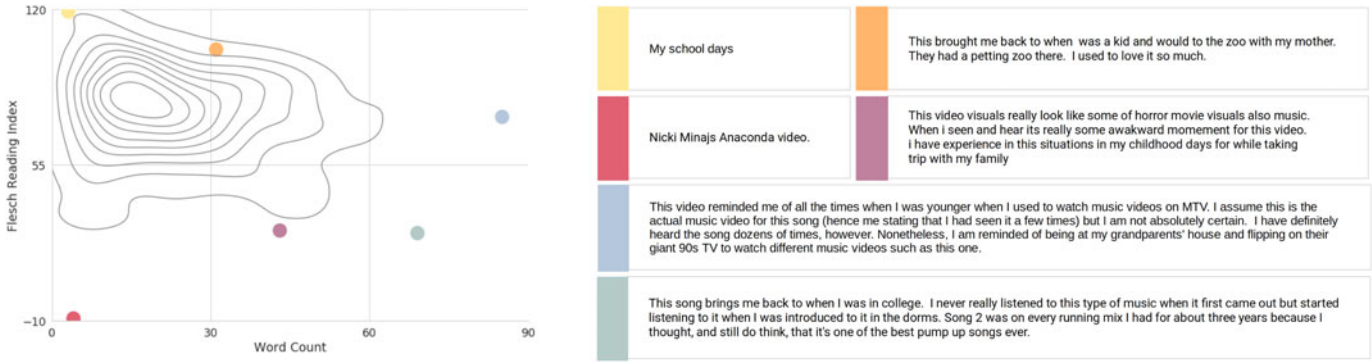

Fig. 4. Contour plot of the Word Count and Flesh Reading-Ease Score (FRES) of the memory descriptions collected from participants. Text fragments on the right are examples from the marked locations.

4.3.1 Visual Data

Duration. Filtering has removed recordings with a large difference in duration from the targeted 60 seconds ($M(SD) = 60.5(2.1)$). The combined duration of all footage captured sums up to a total of 2012 minutes.

Lighting Conditions. Recordings vary broadly in terms of the brightness ($M(SD) = 0.41(0.12)$) and contrast ($M(SD) = 0.21(0.06)$) descriptors (see Fig. 3, for a visual impression of this diversity).

4.3.2 Facial Expressions

OpenFace detected faces successfully as present in most of the frames (99 percent of all available in the dataset) and with a high degree of confidence ($M(SD) = .96(0.07)$). The automatically extracted Action Unit-coding indicates that participants' expressions are subtle: the measure for the average maximal action unit intensity varies across responses around the value of 1, indicating that on average *any* of the coded action units is at most *"present at minimum intensity"* in the OpenFace detections ($M(SD) = 1.22(0.71)$).[5] The overall low rate with which any facial actions are present in a response ($M(SD) = 0.56(0.35)$) further underlines that expressions are likely sparse.

In addition to these quantitative insights, visual inspection of the footage reveals substantial variation in viewers' poses across recordings (e.g., individuals watching videos while laying down on a bed, sometimes with their devices resting on their chest, resulting in camera movements).

5 Analysis of Validity

5.1 Variation and Balance of Affective Ratings

Induced Emotion. A look at the distribution of induced emotion across responses shows that the corpus covers the entire PAD-space (see Fig. 5). However, analysis of the number of responses in the different octants of the 3-dimensional PAD-space reveals a significant imbalance ($\chi^2(7, 1995) = 546.64$, $p < .001$). In particular, there are only a few reports with feelings of *"Anger"* (low in pleasure, high in arousal, and high in dominance) or *"Fear"* (low in pleasure, high in arousal, and low in dominance). While it is plausible that responses to music videos may rarely evoke these kinds of responses in viewers, it is a limitation that users of the corpus should consider for computational modeling (e.g., for facial affect analysis).

Memory-Associated Affect. Similar to induced emotions, ratings for memories span all quadrants in the Pleasure-Arousal and Pleasure-Dominance planes (see Fig. 6). However, further analysis of the distribution of memories over the different octants of the 3-dimensional PAD-space reveals also here substantial imbalances ($\chi^2(7, 989) = 670.24$, $p < .001$). About 60 percent ($N = 606$) of all memories are associated with positive pleasure or dominance, differing only in their arousal. This finding is consistent with empirical data demonstrating a tendency of positive memories to remain more available for recall than negative ones [50]. It might also reflect a bias in the willingness of participants to report negative events in our study. Again, this imbalance is something that should be kept in mind when using the corpus. Consequently, it is prudent to

TABLE 4
Statistics of Behavioral Recordings Collected for Responses to Videos (Processed and Filtered Dataset)

Variable	Measure	$M\ (SD)$	Min/Max
Visual Data	**Duration (Sec.)**	60.5 (2.1)	50.33/69.86
$N = 1995$	**Brightness**	0.41 (0.12)	0.00/0.68
	Contrast	0.21 (0.06)	0.01/0.39
Facial Expr.	**Avg. Max. AU-Int.**	1.22 (0.71)	0.5/4.0
$N = 1995$	**Pres. Facial Actions**	0.56 (0.35)	0/1
	Avg. Conf.	0.96 (0.07)	0.00/0.98

$M(SD)$: *Mean and Standard Deviation;*
Min/Max : *Range of values occurring in the sample;*

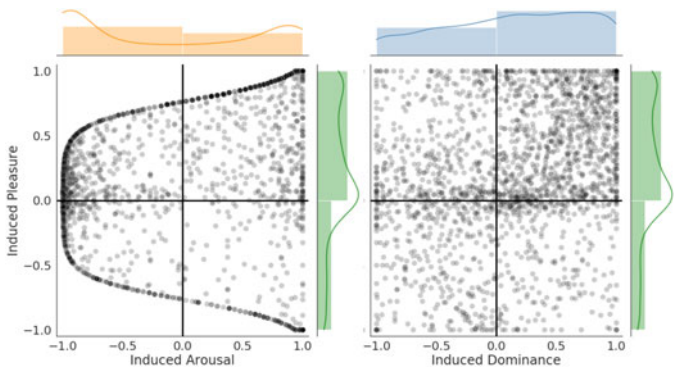

Fig. 5. Distribution of *induced emotion for individual responses* in the pleasure-arousal and pleasure-dominance planes ($N = 1995$).

5. https://github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units

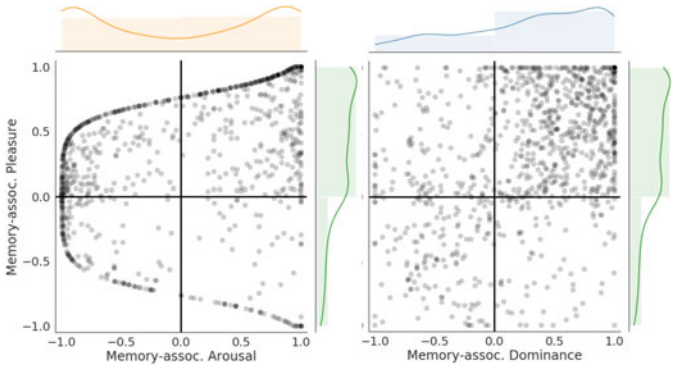

Fig. 6. Distribution of *associated-affect for individual memories* in the pleasure-arousal and pleasure-dominance planes ($N = 989$).

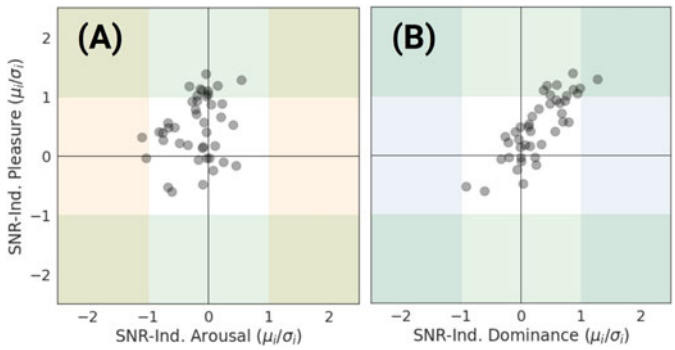

Fig. 7. Plots of the video-wise Signal-to-Noise Ratio (*SNR*) Score for ratings of *Induced Emotions* on *(A)* the pleasure-arousal plane and *(B)* pleasure-dominance plane. Colors demarcate regions where the mean ratings for a video on the respective axis exceed its standard deviation.

use the memories contained in Mementos to primarily study or model differences between neutral and positive associations of pleasure and dominance.

5.2 Effects and Relationships

This section describes the findings of a series of statistical analyses of the self-report measures in Mementos. They demonstrate that the dataset successfully captures different aspects of affect and memory processing in response to videos and underlines how these relate to existing research in psychology. Some of the findings discussed in this section are presented in greater detail in other publications using the Mementos dataset [8], [9], [10].[6]

5.2.1 Induced Emotion

Effect of Video Stimuli. In order to serve as a viable corpus for modeling video-induced emotions, it is important to verify that the stimuli presented to viewers actually had an emotional impact on them. For this purpose, we conduct separate one-way ANOVAs for ratings of Induced Pleasure, Arousal, and Dominance to identify the difference between video stimuli using linear mixed-effects models (DVs: Induced Pleasure, Arousal, or Dominance; IV: Video Identity; Random-Intercept: Participant Identity). Results indicate that there exists statistically significant effects on each dimension of viewers' affective responses (P: $F(41, 1005.26) = 5.57,\ p <= .001,\ R^2_m = .169$); A: $F(41, 969.09) = 4.51,\ p <= .001,\ R^2_m = 0.131$; D:$F(41, 973.63) = 4.55,\ p <= .001,\ R^2_m = 0.129$). However, taken across dimensions, these differences account only for an average of 14 percent of the total variation in responses, leaving the remaining 86 percent unexplained. Consequently, while these findings demonstrate that exposure to the videos does indeed shape viewers' induced emotions, it also suggests that their affective impact independent of context is not very strong. This relationship manifests itself in clear differences in emotional impact among viewers of the same video, i.e., within-video variation. A visual representation demonstrating this phenomenon can be seen in the Fig. 7. It shows the distribution of a video-wise *Signal-to-Noise Ratio (SNR) score*, computed as the ratio of the mean to the standard deviation for its ratings on each affective dimension of PAD-space ($Score_{SNR} = \mu_i/\sigma_i$). The average score of ratings can be considered the effective signal of the stimulus, while variation in responses corresponds to noise distortion. Consequently, a stimulus with an SNR score substantially different from 0 for a particular dimension evokes both pronounced (large $|\mu_i|$) and highly similar responses across different viewers (small σ_i). In particular, values $|Score_{SNR}| >= 1$ indicate that ratings for the stimulus can be considered as *unipolar*. That is, different viewers' responses to the video are sufficiently similar to each other to – on average – not expect responses with an affective polarity opposite to that of their mean. Only responses to some stimuli in the dataset do pass this threshold. In total, 9 videos induce unipolar pleasure, 2 arousal, and only 1 dominance. In summary, these findings demonstrate that videos presented to participants 1) were successful at inducing different emotional responses, and 2) that they differ in the degree of within video-variation that they elicit. These are both properties that we aimed for when designing the corpus, strengthening its validity as a resource for modeling emotional responses to videos. Furthermore, despite their limited number, video stimuli with unipolar responses might be useful for targeted emotion induction procedures in experiments.

Influence of Personal Memories. A detailed analysis of the influence of personal memories on video-induced emotions for the responses in Mementos, as well as a discussion of its relevance for developing context-sensitive automated predictions, can be found in Dudzik *et al.* [8]. Principal findings include that *(1)* responses to videos involving the recollection of memories are associated with higher average levels of induced pleasure, arousal and dominance compared to responses that do not, and that *(2)* that ratings of memory-associated affect are strong predictors of video induced emotions. These findings are in line with earlier empirical work investigating this relationship to media content [13], [51]. Overall, they point to the validity of Mementos as a corpus capturing interactions between personal memories and affective processing.

Individual Differences and Mood-Effects. When controlling for the influence of personal memories, the viewer-specific measures captured in the dataset provide only negligible insights into induced pleasure, arousal, and dominance (see Dudzik *et al.* [8] for the detailed analysis). In particular, we find that viewers' personality does not have a significant effect on their induced emotions under these circumstances,

6. Note that analyses in these publications are based on slightly differently curated versions of the dataset, i.e., without filtering data for multimodal completeness.

and differences in demographics and mood only have a small impact (Demographics: $Avg\Delta R_m^2 = .013$, Mood: $Avg\Delta R_m^2 = .014$). This weak performance underlines the overall difficulty of accounting for variation in emotional responses and the potential of exploiting information about relevant personal memories for improving automated predictions. However, in the absence of such memory information, viewer-specific measures do still offer valuable insights. Analysis with separate linear-mixed effects regressions shows that using all of the viewer-specific measures in Mementos together as predictors for responses without recollections (DVs: Induced Pleasure, Arousal, or Dominance; IVs: Demographics, Personality Scores, and Mood;) accounts for an average of 5 percent of the variance across induced pleasure ($F(13, 223.33) = 2.14, p < .05, R_m^2 = 0.032$), arousal ($F(13, 232.87) = 3.12$, $p < .001$, $R_m^2 = 0.06$) and dominance ($F(13, 232.48) = 2.94$, $p < .001$, $R_m^2 = 0.057$). Together, this shows that Mementos captures individual differences and mood effects, mirroring findings in other research on responses to video content (e.g., [7], [33]), thereby adding to the validity of the corpus.

5.2.2 Memory Evocation

Effects of Video Stimuli and Familiarity. Previous findings indicate that video stimuli can substantially differ in their capacity to trigger personal memories [11]. In particular, for musical material, one variable associated with its evocative potential is familiarity with it [12]. As such, we expect stimuli in our dataset to differ in their capacity to trigger personal memories, which should depend on viewers' familiarity with them. To explore whether these effects are present in our dataset, we use a mixed-effect logistic regression to model the probability of a response to involve any memories, i.e., at least one (DV: Recollection; IV: Video Identity; Random-Intercept: Participant Identity). Results show a statistically significant effect ($\chi^2(31, 1995) = 78.43$, $p < .001$), indicating that the videos in Mementos systematically differ in their evocative potential (see Fig. 8 A for an illustration of the video-wise differences in the rate at which videos evoked recollections). To explore the influence of familiarity, we expand this model by including the effects of viewers' previous exposure (DV: Recollection; IV: Video Identity, Familiarity, and 2-way interaction; Random-Intercept: Participant Identity). Separate likelihood-ratio tests for each effect indicate that only previous exposure remains as a statistically significant effect ($\chi^2(31, 1995) = 78.43, p < .001$). These findings suggest that viewers' familiarity with the material fully mediates differences in videos' capacity to trigger memories (see Fig. 8 B for a visualization of this relationship). Both the differences in evocativeness and the role of familiarity are consistent with existing research, further indicating the validity of Mementos.

Influence of Age-Differences. As part of designing the data collection procedure, we constrained participants' age to a range for which we expected it likely that they would have associated personal experiences that our videos can trigger. Consequently, because we constrained variation in age as part of our data collection design, we would expect it to play no systematic role in the occurrence of recollections. Nevertheless, analysis with a mixed-effects logistic regression (DV: Recollection; IV: Age; Random-Intercepts: Participant and Video Identity), reveal a weak, but statistically significant effect of increased age on the occurrence of recollection ($\beta = 0.23$; $SE = 0.10$; $z = 2.34, p < .05$). This result indicates that we likely could have triggered more memories with our set of videos by constraining our sample of participants to a slightly higher age range. However, it also provides tentative evidence for congruence with established findings on the role of age in memory retrieval that we tried exploiting in our design to maximize triggered memories.

5.2.3 Memory-Associated Affect

Influence of Vividness. Findings from empirical psychology indicate that the clarity and vividness with which memories are recollected is proportional to the intensity of the emotional meaning attributed to them [52]. An analysis of this relationship in our dataset with a mixed-effects regression (DV: Memory-associated Affect Intensity; IV: Vividness; Random-Intercepts: Video and Participant Identity) reveals a weak, but statistically significant correlation ($\beta = 0.22$, $SE = 0.03$, $t(843.84) = 6.78$, $p < .001$). Moreover, regressions of vividness scores on the word count measure for free-text memory descriptions (DV: Vividness; IV: Word Count; Random-Intercepts: Video and Participant Identity) also indicate that viewers tend to describe vivid memories in greater detail ($\beta = 0.133$, $SE = 0.037$, $t(958.35) = 3.97$, $p < .001$). Together, this demonstrates relationships consistent with existing research and provides evidence for the validity of free-text as a potential resource for modeling memory experience.

Mood-Congruent Recall. Mood has been identified as an important influence shaping memories that individuals recollect, a phenomenon referred to as *Mood-congruent recall* [53]. We conducted regression analyses to identify whether mood primes memory-associated affect in our dataset (DV: Memory-associated Pleasure, Arousal or Dominance (either); IVs: Mood Pleasure, Arousal and Dominance (all); Random-Intercepts: Participant and Video Identity). Results reveal weak partial correlations between matching affective dimensions: mood pleasure is positively correlated with memory pleasure ($\beta = 0.18$, $SE = 0.04$, $t(223.55) = 4.07$, $p < .001$), mood arousal with memory arousal ($\beta = 0.12$, $SE = 0.05$, $t(207.54) = 2.45, p < .05$), and mood dominance with memory dominance ($\beta = 0.12$, $SE = 0.04$, $t(224.58) = 3.44$, $p < .001$). However, with an average explained variance of 2.5 percent across models, the overall effect of this mood-concurrency is comparatively weak. This finding indicates that recollections in Mementos are subject to mild mood-congruent priming effects and that considering these might benefit modeling memory-associated affect.

5.3 Analysis of Multimodal Data

5.3.1 Webcam Recordings

Impact of Lighting Conditions on Facial Analysis. Lighting conditions can pose a challenge for vision-based face analysis [54]. The recordings of faces in Mementos vary substantially in their brightness and contrast, reflecting whatever environment viewers chose to participate in. To understand the potential impact of lighting conditions in Mementos, we conduct a regression analysis of these factors on the average

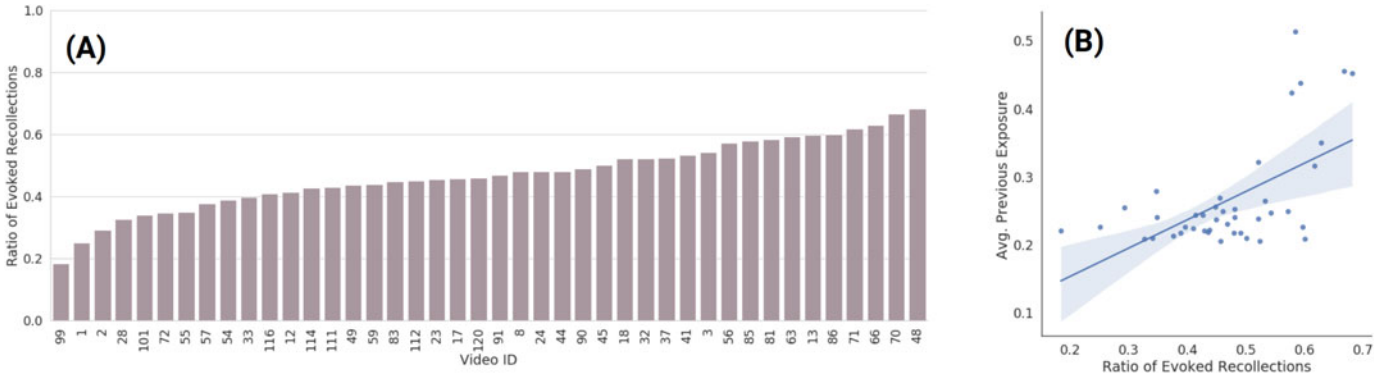

Fig. 8. (A) Video-wise rates at which exposure evoked a recollection (i.e., at least one personal memory is triggered). (B) Scatter plot with linear relationship between the video-wise Rates of Evoked Recollections and Viewers' Average Degree of Previous Exposure to them. Shaded area denotes the 95 percent confidence interval.

confidence with which OpenFace detects faces in a recording (DV: Confidence; IV: Brightness and Contrast; Random-Intercepts: Video and Participant Identity). This reveals statistically significant effects of the brightness ($\beta = -0.24$, $SE = 0.05$, $t(841.58) = -4.92$, $p < .001$) and contrast ($\beta = 0.21$, $SE = 0.05$, $t(1057.87) = 4.70$, $p < .001$). However, the magnitude of these effects is small, and the overall confidence scores for automatic analysis of OpenFace are both high and fairly stable ($(M(SD) = 0.96(0.07)$). Consequently, this indicates that differences in lighting conditions are a potential limitation and can impact automatic analysis. As such, they should be kept in mind when using the dataset for automatic behavioral analysis, even though the overall impact of these conditions on state-of-the-art approaches is likely negligible.

Differences in Lighting Conditions Across Induced Emotions. Given the effect of lighting conditions on automatic analysis, we further investigate whether there are systematic differences across different PAD space regions. Such imbalances would be undesirable since they may negatively bias the performance of automatic analyses. For this purpose, we conduct separate one-way ANOVAs using mixed linear regression models (DV: Brightness or Contrast; IV: PAD-Octant; Random-Intercepts: Participant and Video Identity). Results reveal no statistically significant differences between the mean brightness or contrast across the octants of PAD-space. This finding suggests that any influence of lighting conditions will not be systematically impacting particular types of affective responses.

5.3.2 Free-Text Memory Descriptions

Differences in Text Complexity Across Associated Affect. It is plausible that people may express memories with certain affective associations less detailed than others (e.g., when connected to negative feelings of sadness or fear). Since these differences might be relevant for computational analysis, we investigate whether our measures for text complexity differ across the octants of the PAD space. An analysis with separate one-way ANOVAs using mixed linear regression models (DV: Word Count or Contrast; IV: PAD-Octant; Random-Intercepts: Participant and Video Identity) reveals no significant differences across octants for either the Word Count or the FRES metric. This finding indicates that the corpus contains memory descriptions with a similar level of detail across the entire PAD space.

Correspondence of Human Interpretations. We have previously explored the capacity for human readers to correctly infer affective meaning from the free-text memory descriptions in Mementos [9]. We summarize these efforts here, since their findings can give insights into the potential performance of computational approaches on the data. We let two annotators rate pleasure, arousal and dominance for a selection of 150 memory descriptions (140 of which remain in the curated dataset) in terms of their 1) *Perceived Conveyed Affect (PCA)*, and 2) *Inferred Affective Experience (IAX)* of the author. For PCA ratings, readers respond to the question *"What feelings does this text express?"*, and were instructed only to consider explicitly expressed affect, e.g., emotion words. Performance on this task provides insights into how explicit authors describe their emotions in the text. In the case of IAX ratings, annotators answer the question *"How do you think the person describing this memory feels about it? Put yourself into their situation"*. The motivation for this different task formulation is to encourage annotators to use their own knowledge and experience to infer implicit emotional meaning (e.g., by drawing on stereotypical affective meaning of event memories, such as weddings or parties). Findings revealed that raters' judgments in both tasks for pleasure and dominance moderately correlated with self-reported memory-associated affect. However, correspondence dropped substantially for arousal. Similarly, the average degree of correspondence across affective dimensions was greater for the IAX task than the PCA task. Together, these findings indicate that the free-text memory descriptions contain information enabling human readers to relate to how viewers felt about their memories, but that doing so might be particularly challenging for arousal. Moreover, given the stark differences in raters' performance between the IAX and the PCA tasks for arousal, a potential reason for this might have been a lack of explicit expressions (i.e., arousal-related emotion words). Consequently, it may be challenging for automatic approaches that rely on such expressions to make accurate inferences. This conclusion is further supported by findings from our own prior experiments in which we extracted a broad range of affective lexical features from descriptions to predict induced emotions [9], [10] (see also Section 6 below).

Confidence of Human Interpretations. Alongside the PCA and IAX affect ratings collected from our two readers, we asked them to also indicate their confidence when doing so (9-point Likert Scale; 1-totally uncertain to 9-very certain). We have not reported on this data previously and do so here for the sample of 140 descriptions that remain in the curated dataset. Their analysis can provide additional insights into the ease of human interpretation of free-text descriptions, and thus their potential for automatic analyses. Results of a correlational analysis show only a moderate agreement between our two readers' confidence on the PCA task ($r(138) = .372,\ p < .001$),[7] and none between their more subjective IAX ratings. However, pair-wise averaging of these ratings suggests an overall high degree of confidence (PCA: $M(SD) = 7.43(1.08)$; IAX:$M(SD) = 6.38(1.1)$).

Impact of Text Complexity on Human Affective Interpretation. To explore the impact of memory descriptions' text complexity on the ease with which they can be emotionally interpreted, we analyze its relation to the confidence and error of our human readers' affective ratings. Additionally, we also look at its relation to the absolute errors these raters made for pleasure, arousal, and dominance when guessing memory-associated affect. Regression analyses (DV: Confidence; IV: FRES and Word Count) shows no significant partial correlations between either measure for memory descriptions with annotators' pair-wise averaged confidence ratings for either PCA or IAX.[7] Similarly, separate regression analyses for the absolute error our raters' guesses for pleasure, arousal, and dominance (DV: Abs. Error; IV: FRES and Word Count) reveal no significant relationships.[7] These findings suggest that the detail of descriptions – as quantified by our measures – has no adverse effect on the error or confidence of our raters. Since this is the case for both the PCA and the IAX task, it seems plausible to expect no adverse effects of the text complexity of descriptions on automatic analyses as well.

6 Evidence for Usefulness

This section reports on a series of studies that have successfully used the multimodal data in Mementos for machine learning experiments on automatic affect prediction. In particular, they provide salient examples for the types of research questions that can be addressed with the dataset and serve as baseline approaches for doing so.

6.1 Context-Sensitive Video Affective Content Analysis

Traditionally, approaches for Video Affective Content Analysis (VACA) do not address within-video variation by incorporating information about viewers' context. Using Mementos, we have previously explored a multimodal approach that leverages memory descriptions as context for VACA and compares it to a context-independent approach [9]. Concretely, we extracted distinct feature sets to represent videos' audiovisual content and the free-text descriptions of viewers' personal memories. Using an ablation study setup, we then explored the performance achieved by these different modalities for predicting the affect induced in individual viewers. For this purpose, we compared the performance between two different approaches: one using feature-level fusion (concatenation of modality-specific features with a support vector regressor for prediction) and another using late-fusion (training of separate modality-specific models combined via stacked generalization for prediction with an L2-regularized linear model as meta regressor).

Our experiments demonstrate that analyzing viewers' memory content in addition to videos' audiovisual content provides substantial information about within-video variation, especially for induced pleasure and dominance. In comparison, arousal performed relatively poorly. Further investigation of memory descriptions with data collected from human annotators reveals a similar pattern in performance (see Section 5.3.2). Notably, our approach using only video features already performed similarly to a perfect oracle for context-free VACA, i.e., a model that always predicts the accurate video-wise average for induced pleasure, arousal, and dominance. Finally, our comparison between early- and late-fusion revealed better performance for the latter. This shows the potential of this fusion approach in multimodal modeling for this task over simple feature-concatenation, despite the increased complexity of implementation.

6.2 Use for Affective Behavior Analysis

Automatic approaches for affect detection often use facial analysis in isolation, without incorporating additional aspects of the wider context. This way of inferring affect is strikingly different from how human perceivers make sense of behavioral signals [55] and limits performance in real-world scenarios. For this reason, we have explored the potential of automatically analyzing video and memory content alongside facial behavior to support affect detection in a previous study, using the Mementos dataset [10]. Besides extracting distinct feature sets for representing video content and descriptions of viewers' personal memories, we used OpenFace to analyze their facial behavior (Action Units, Eye Gaze-, and Head Pose-features). Our approach for predictive modeling consisted of an array of modality-specific support vector regressors combined via late-fusion with a meta regressor (L2-regularized linear model, stacked generalization). Using an ablation study setup for our experiments, we then explored the contribution of both context modalities next to facial analysis on affect prediction performance. While our findings confirmed that adding context provides overall performance improvements, they also offer insights into the complementary nature of affective information sources. Notably, facial expressions provided unique benefits for predicting arousal, while video and memory content explained unique variation in viewers' pleasure and dominance. Together, this study highlights both the potential performance benefits of context-sensitive predictions for real-world applications, as well as the possibility of intelligent trade-offs for predicting particular aspects induced emotions. Importantly, it shows both 1) the suitability of the behavioural recordings in Mementos for modelling viewers' experienced affect, as well as 2) the challenges that this approach faces in the realistic setting it captures: uncontrollable recording

7. Statistical significance tested using clustered bootstrapping ($B = 10000$ repetitions) to account for the nesting of memory descriptions in participants.

conditions and potentially sparse facial expressiveness from participants.

7 Potential Uses in Future Research

7.1 Modeling Video-Induced Emotion

The primary use of Mementos is as a resource for developing and testing computational models of video-induced emotions. In particular, this encompasses the two types of research strains addressed in the studies discussed in Section 6: detecting affect by analyzing audiovisual recordings of individuals' (non-verbal) behaviors (i.e., *Affect Detection*), or the automatic analysis of the audiovisual content of consumed videos (i.e., *Video Affective Content Analysis (VACA)*).

As a primarily viewer-centric corpus (see our discussion in Section 2.4), Mementos is particularly suited for affect detection, offering a high degree of ecological validity in terms of recording conditions and emotional responses (experience and behavior) while still providing a substantial amount of data for development purposes. Moreover, because it captures a broad range of contextual factors, it is well suited for work on context-sensitive approaches for affective analysis. Researchers could use Mementos to hone in on the influence of personal memories as we have done in our prior research with the corpus [10], or they may focus more extensively on the other contextual factors it contains, such as Personality or Mood. The comparatively small amount of unique video stimuli and their limited diversity (i.e., only music videos) make it likely not suitable for traditional VACA research, which is generally interested in modeling coarse differences in affective impact *across* videos [2]. Such research would require a more video-centric corpus. Nevertheless, our analyses of Mementos demonstrate that it contains substantial information on the within-video variation of affective responses and contextual factors connected to it. As such, it forms a valid and valuable resource for further explorations on the topic of personalized and context-sensitive VACA approaches (see the reviews of Wang *et al.* [2], and Baveye *et al.* for the importance of research on context-sensitivity in VACA [4], as well as Soleymani *et al.* [7] for a comprehensive discussion of addressing context in corpora for VACA research). Alternatively, we encourage future work to extend – or build on –, Mementos to construct a video-centric corpus for context-sensitive traditional VACA research (e.g., by collecting additional data on other types of video content).

7.2 Modeling Memory-Associated Affect

Independent of their influence on video experiences, modeling personal memories' emotional interpretation is a worthwhile goal in its own right. People's evaluation of past moments is a crucial influence on their intentions for the future [56]. In particular, information about the affect associated with past experiences involving products and services might be important for designing and personalizing these [57]. Similarly, analysis of written reflections about the past might benefit the development of technology for supporting psycho-social well-being, e.g., by negative memories as a potential symptom for depression [58]. Mementos contains examples of how people describe their memories using free text. As such, researchers could use Mementos as a resource for affective text analysis or sentiment mining to that end. Moreover, people likely display individual differences and culture-specific ways of expressing their memory-associated affect in text. The viewer-specific measures in Mementos enable researchers to explore such potential influences. Finally, the findings of our preliminary analyses and previous computational work that we have presented in this article point to a challenge for understanding memory descriptions solely based on explicit expressions of affect contained in them – particularly for arousal. Future research could use Mementos as a resource to develop and test technical approaches for inferring the necessary implicit affective meaning.

7.3 Modeling Memory Evocation

Detecting attentional shifts away from externally located stimuli towards mental content during peoples' interactions with technology or media content is an emerging field of research. A primary reason for this is that awareness of such mind wandering under the wrong circumstances can potentially avoid negative (e.g., unfocused students [59]) or even disastrous consequences (e.g., a distracted driver on the road [60]). The recollection of personal memories can require significant internal attentional engagement [61]. As such, these may result in behavioral responses similar to other types of internal cognitive processing, which can be detected through gaze behavior [62]. Modeling this relationship is in principle feasible from audiovisual data, and as such, the recordings in Mementos might serve as a resource for exploring technological approaches under real-world conditions. Moreover, given the potential influence of personal memories on viewers' emotional processing of video content, research on affect-adaptive media technology might be interested in modeling the evocativeness of video content. This is especially true since viewers differ in how they experience a stimulus with or without associated memories [8]. Consequently, while the amount of unique videos used in Mementos is limited, it might nevertheless provide a unique starting point for such exploration. Moreover, the existence of context-effects for evocativeness invites researchers to investigate these influences in such modeling activities.

8 Limitations

Despite striving to maximize the ecological validity of the collected data, there are limitations to how the captured responses may generalize to other types of media material or a different viewership. First, the dataset considers only responses to a particular media content format, i.e., music videos. It is plausible that other types of material may result in different emotional responses and be subject to a different influence of personal memories (e.g., feature films, where empathy with protagonists in the narrative is an important mechanism [63]). Moreover, even the music videos in the corpus are only of limited variety in terms of genre (i.e., mainly variants of Rock and Pop) and release years (i.e., the 2000s). Consequently, these are likely not representative of the wider populations' musical preferences. Importantly, we selected stimuli to maximize the chance for memories to occur and influence responses to content. As such, our data

is likely not representative of how often memories occur and influence responses in the general population or for other types of video content. Future work could create corpora to measure the occurrence of memories in even less constrained settings and involving a greater diversity of media content.

Similarly, limitations may apply to the shape and form of the free-text memory descriptions that participants' provided. Despite striving for ecological validity, our setting is still taking place in an online survey context. Consequently, the reported free-text descriptions might differ from how people would report on their memories in a social media setting or writing in a diary. Future research might attempt to collect such descriptions from comments from videos on social media or using a methodology similar to the reminiscence application described in Peesapati *et al.* [64]

Another set of limitations of Mementos as a corpus for predictive modeling is the imbalanced distribution of examples. First, responses mostly involve positive or neutral affect, both for induced emotions and memory-associated affect. Similarly, the dataset contains a substantial imbalance in participants' nationalities and personalities, likely reflecting the distribution of users on Mechanical Turk. The development of future corpora that explicitly balances nationality (e.g., similar to SEWA [30]) and personality in recruitment may improve this. Presently, however, these imbalances are something that researchers should be aware of when relying on the dataset and address them where appropriate (e.g., by specialized sampling procedures for training classifiers on imbalanced data [65], or using relevant subsets of the corpus).

Finally, our protocol only revolved around personal memories in isolation. We did not try to capture any other types of mental responses that might have occurred, e.g., semantic associations. Future data collection efforts could improve this and explicitly code for different mental responses (e.g., based on the scheme in a study by McDonald *et al.* [11]). This could be combined with a fine-grained classification scheme for memory content (e.g., according to types of life events, similar to the one deployed by Nazareth *et al.* [66]).

9 Conclusion

In this paper, we have presented the first multimodal dataset capturing the occurrence and influence of personal memories on affective responses to video stimuli in-the-wild. We have argued for its validity as a dataset for computational modeling by providing evidence for the diversity of affective responses covered by it, its congruence with existing findings from psychology about affect and memory processing, and an analysis of its multimodal data.

Because of 1) its range of relevant content (self-report measures, free-text memory descriptions and behavioural recordings), and 2) its high degree of ecological validity, Mementos lends itself as a valid resource for future computational research on *Video-Induced Emotions*, *Memory-Associated Affect*, and *Memory Evocation*. This article has reviewed the two existing studies in which we have previously relied on Mementos for multimodal machine learning experiments. While they demonstrate the corpus' principal usefulness for multimodal modeling, our investigations have solely touched upon the first of these three research topics. Consequently, we encourage using Mementos for future work in line with our own and as a readily available resource for modeling these two alternative – and largely unexplored – aspects of human affect and memory processing.

Acknowledgments

Authors would like to thank Santosh Ilamparuthi for his support in developing the procedures for data collection and management and Tim Rietveld for contributing to the data analysis. This work was supported by the 4TU Research Center Humans & Technology (H&T) Project (Systems for Smart Social Spaces for Living Well: S4).

References

[1] A. Bartsch, "Emotional gratification in entertainment experience. Why viewers of movies and television series find it rewarding to experience emotions," *Media Psychol.*, vol. 15, no. 3, pp. 267–302, Jul. 2012.

[2] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 410–430, Oct.–Dec. 2015.

[3] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, Mar. 2006.

[4] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 396–409, Oct.–Dec. 2018.

[5] A. Stewart, N. Bosch, H. Chen, P. Donnelly, and S. D'Mello, "Face forward: Detecting mind wandering from video during narrative film comprehension," in *Proc. Int. Conf. Artif. Intell. Educ., 2017*, pp. 359–370.

[6] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, and A. Oliva, "Multimodal memorability: Modeling Effects of semantics and decay on video memorability," Sep. 2020, *arXiv:2009.02568*.

[7] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1075–1089, Jun. 2014.

[8] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, "Investigating the influence of personal memories on video-induced emotions," in *Proc. 28th ACM Conf. User Model. Adapt. Personalizat.*, Jul. 2020, pp. 53–61.

[9] B. Dudzik, J. Broekens, M. Neerincx, and H. Hung, "A blast from the past: Personalizing predictions of video-induced emotions using personal memories as context," 2020, *arXiv:2008.12096*.

[10] B. Dudzik, J. Broekens, M. Neerincx, and H. Hung, "Exploring personal memories and video content as context for facial behavior in predictions of video-induced emotions," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, vol. 10, pp. 153–162.

[11] D. G. McDonald, M. A. Sarge, S.-F. Lin, J. G. Collier, and B. Potocki, "A role for the self," *Commun. Res.*, vol. 42, no. 1, pp. 3–29, Feb. 2015.

[12] P. Janata, S. T. Tomic, and S. K. Rakowski, "Characterisation of music-evoked autobiographical memories," *Memory*, vol. 15, no. 8, pp. 845–860, Nov. 2007.

[13] H. Baumgartner, M. Sujan, and J. R. Bettman, "Autobiographical memories, affect, and consumer information processing," *J. Consum. Psychol.*, vol. 1, no. 1, pp. 53–82, Jan. 1992.

[14] E. van den Hoven, "A future-proof past: Designing for remembering experiences," *Memory Stud.*, vol. 7, no. 3, pp. 370–384, Jul. 2014.

[15] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, "Artificial empathic memory," in *Proc. Workshop Understanding Subjective Attributes Data Focus Evoked Emot.*, 2018, pp. 1–8.

[16] C. Lagger, M. Lux, and O. Marques, "What makes people watch online videos," *Comput. Entertainment*, vol. 15, no. 2, pp. 1–31, Apr. 2017.

[17] K. H. Greenaway, E. K. Kalokerinos, and L. A. Williams, "Context is everything (in emotion research)," *Soc. Pers. Psychol. Compass*, vol. 12, no. 6, Jun. 2018, Art. no. e12393.

[18] S. M. Smith and E. Vela, "Environmental context-dependent memory: A review and meta-analysis," *Psychon. Bull. Rev.*, vol. 8, no. 2, pp. 203–220, Jun. 2001.
[19] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 43–55, Jan.–Mar. 2015.
[20] D. McDuff and M. Soleymani, "Large-scale affective content analysis: combining media content features and facial reactions," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.* May 2017, pp. 339–345.
[21] J. J. Sun, T. Liu, A. S. Cowen, F. Schroff, H. Adam, and G. Prasad, "EEV Dataset: Predicting expressions evoked by diverse videos," Jan. 2020, *arXiv:2001.05488*.
[22] S. Carvalho, J. Leite, S. Galdo-Álvarez, and Ó. F. Gonçalves, "The emotional movie database (EMDB): A self-report and psychophysiological study," *Appl. Psychophysiol. Biofeedback*, vol. 37, no. 4, pp. 279–294, Dec. 2012.
[23] S. Koelstra *et al.*, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.
[24] J. A. MirandaCorrea, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr.–Jun. 2021.
[25] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Apr.–Jun. 2018.
[26] S. C. Guntuku, M. J. Scott, H. Yang, G. Ghinea, and W. Lin, "The CP-QAE-I: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia," in *Proc. Seventh Int. Workshop Qual. Multimedia Experience*, May 2015, pp. 1–7.
[27] 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 Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Jul.–Sep. 2015.
[28] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.
[29] Y.-G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *Proc. Twenty-Eighth AAAI Conf. Artif. Intell.*, Jun. 2014, pp. 73–79.
[30] J. Kossaifi *et al.*, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021.
[31] B. Dudzik *et al.*, "Context in human emotion perception for automatic affect detection: A survey of audiovisual databases," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, Sep. 2019, pp. 206–212.
[32] L. M. Jenkins and D. G. Andrewes, "A new set of standardised verbal and non-verbal contemporary film stimuli for the elicitation of emotions," *Brain Impairment*, vol. 13, no. 2, pp. 212–227, sep. 2012.
[33] S. C. Guntuku, W. Lin, M. J. Scott, and G. Ghinea, "Modelling the influence of personality and culture on affect and enjoyment in multimedia," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, Sep. 2015, pp. 236–242.
[34] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Sci. Inf.*, vol. 44, no. 4, pp. 695–729, Dec. 2005.
[35] J. Broekens, T. Bosse, and S. C. Marsella, "Challenges in computational modeling of affective processes," *IEEE Trans. Affect. Comput.*, vol. 4, no. 3, pp. 242–245, Jul.–Sep. 2013.
[36] J. A. Allen *et al.*, "Comparing social science and computer science workflow processes for studying group interactions," *Small Group Res.*, vol. 48, no. 5, pp. 568–590, Oct. 2017.
[37] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychol.*, vol. 14, no. 4, pp. 261–292, Dec. 1996.
[38] J. Broekens and W.-P. Brinkman, "AffectButton: A method for reliable and valid affective self-report," *Int. J. Hum. Comput. Stud.*, vol. 71, no. 6, pp. 641–667, 2013.
[39] P. J. Lang, M. M. Bradley, B. N. Cuthbert, and Others, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center Study Emotion Attention*, vol. 1, pp. 39–58, 1997.
[40] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–36, Feb. 2015.
[41] R. Samide, R. A. Cooper, and M. Ritchey, "A database of news videos for investigating the dynamics of emotion and memory," *Behav. Res. Methods*, vol. 52, no. 4, pp. 1469–1479, Aug. 2020.
[42] R. Cohendet, C.-H. Demarty, Q. K. N. Duong, M. Sjöberg, B. Ionescu, and T.-T. Do, "MediaEval 2018: Predicting media memorability task," in *Proc. MediaEval Workshop*, 2018, pp. 1–3.
[43] M. A. Conway and S. Haque, "Overshadowing the reminiscence bump: Memories of a struggle for independence," *J. Adult Develop.*, vol. 6, no. 1, pp. 35–44, 1999.
[44] R. E. De Vries, "The 24-item brief HEXACO inventory (BHI)," *J. Res. Pers.*, vol. 47, no. 6, pp. 871–880, Dec. 2013.
[45] J. R. de Leeuw, "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," *Behav. Res. Methods*, vol. 47, no. 1, pp. 1–12, Mar. 2015.
[46] R. Reisenzein, "Pleasure-arousal theory and the intensity of emotions," *J. Pers. Social Psychol.*, vol. 67, no. 3, pp. 525–539, 1994.
[47] R. Flesch, "A new readability yardstick," *J. Appl. Psychol.*, vol. 32, no. 3, pp. 221–233, Jun. 1948.
[48] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 59–66.
[49] Z. Cao, G. HidalgoMartinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[50] W. R. Walker, R. J. Vogl, and C. P. Thompson, "Autobiographical memory: Unpleasantness fades faster than pleasantness over time," *Appl. Cogn. Psychol.*, vol. 11, no. 5, pp. 399–413, Oct. 1997.
[51] H. Baumgartner, "Remembrance of things past: Music, autobiographical memory, and emotion," *Adv. Consum. Res.*, vol. 19, pp. 613–620, 1992.
[52] J. M. Talarico, K. S. LaBar, and D. C. Rubin, "Emotional intensity predicts autobiographical memory experience," *Memory Cognit.*, vol. 32, no. 7, pp. 1118–1132, Oct. 2004.
[53] G. E. Matt, C. Vázquez, and W. K. Campbell, "Mood-congruent recall of affectively toned stimuli: A meta-analytic review," *Clin. Psychol. Rev.*, vol. 12, no. 2, pp. 227–255, Jan. 1992.
[54] R. Gopalan and D. Jacobs, "Comparing and combining lighting insensitive approaches for face recognition," *Comput. Vis. Image Understanding*, vol. 114, no. 1, pp. 135–145, Jan. 2010.
[55] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Interest*, vol. 20, no. 1, pp. 1–68, Jul. 2019.
[56] D. Wirtz, J. Kruger, C. N. Scollon, and E. Diener, "What to do on spring break?," *Psychol. Sci.*, vol. 14, no. 5, pp. 520–524, Sep. 2003.
[57] D. A. Norman, "The way I SEE ITMemory is more important than actuality," *Interactions*, vol. 16, no. 2, pp. 24–26, Mar. 2009.
[58] S. Mihailova and L. Jobson, "Association between intrusive negative autobiographical memories and depression: A meta-analytic investigation," *Clin. Psychol. Psychother.*, vol. 25, no. 4, pp. 509–524, Jul. 2018.
[59] F. Putze, D. Küster, S. Annerer-Walcher, and M. Benedek, "Dozing off or thinking hard?," in *Proc. Int. Conf. Multimodal Inter.*, 2018, pp. 258–262.
[60] C. L. Baldwin, D. M. Roberts, D. Barragan, J. D. Lee, N. Lerner, and J. S. Higgins, "Detecting and quantifying mind wandering during simulated driving," *Front. Hum. Neurosc.*, vol. 11, Aug. 2017, Art. no. 406.
[61] M. M. Chun, J. D. Golomb, and N. B. Turk-Browne, "A Taxonomy of external and internal attention," *Annu. Rev. Psychol.*, vol. 62, pp. 73–101, Jan. 2011.
[62] M. Benedek, R. Stoiser, S. Walcher, and C. Körner, "Eye behavior associated with internally versus externally directed cognition," *Front. Psychol.*, vol. 8, Jun. 2017, Art. no. 1092.
[63] J. J. Igartua, "Identification with characters and narrative persuasion through fictional feature films," *Communications*, vol. 35, no. 4, pp. 347–373, Jan. 2010.
[64] S. T. Peesapati, V. Schwanda, J. Schultz, M. Lepage, S.-Y. Jeong, and D. Cosley, "Pensieve," in *Proc. 28th Int. Conf. Hum. Factors Comput. Syst.*, 2010, pp. 2027–2036.

[65] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[66] D. S. Nazareth, M.-P. Jansen, K. P. Truong, G. J. Westerhof, and D. Heylen, "MEMOA: Introducing the multi-modal emotional memories of older adults database," in *Proc. 8th Int. Conf. Affective Comput. Intell. Interact.*, Sep. 2019, pp. 697–703.

Bernd Dudzik (Student Member, IEEE) received the bachelor's degree in online media from Furtwangen University, Germany, and the master's degree in media technology from Leiden University. He is currently working toward the doctoral degree with the Intelligent Systems Department, Delft University of Technology. His current research focuses on building computational models of when and how personal memories influence individuals' emotional experiences of media content. He is currently a member of the Association for the Advancement of Affective Computing.

Hayley Hung (Member, IEEE) received the PhD degree in computer vision from the Queen Mary University of London in 2007. She is currently an associate professor with the Pattern Recognition and Bio Informatics Group, Delft University of Technology and the head of the Socially Perceptive Computing Lab. Between 2010 to 2013, she held a Marie Curie fellowship with the Intelligent Systems Lab, University of Amsterdam. From 2007 to 2010, she was a postdoctoral researcher with Idiap Research Institute, Switzerland. Her research interests include social signal processing, computer vision, and machine learning.

Mark Neerincx received the PhD degree in psychology from the University of Groningen, The Netherlands. He is currently a principal scientist with TNO and a professor of human-centered computing with TU Delft, The Netherlands. His research interests include cognitive engineering, social robots, and cognitive task load modeling for adaptive interfaces.

Joost Broekens is currently the president-elect of the Association for the Advancement of Affective Computing, an associate professor and head of the Affective Computing and Human-Robot Interaction Group at the Leiden Institute of Advanced Computer Science, Leiden University. He is also co-founder and CTO of Interactive Robotics. His research interests focus on affective computing, particularly the computational modeling of emotions and intelligent interaction between humans and socially interactive agents. He is currently an associate editor for the *IEEE Transactions on Affective Computing* and *Sage Adaptive Behavior journal*.