Towards Artificial Empathic Memory

Accounting for the Influence of Personal Memories in Automatic Predictions of Affect

Dudzik, B.J.W.

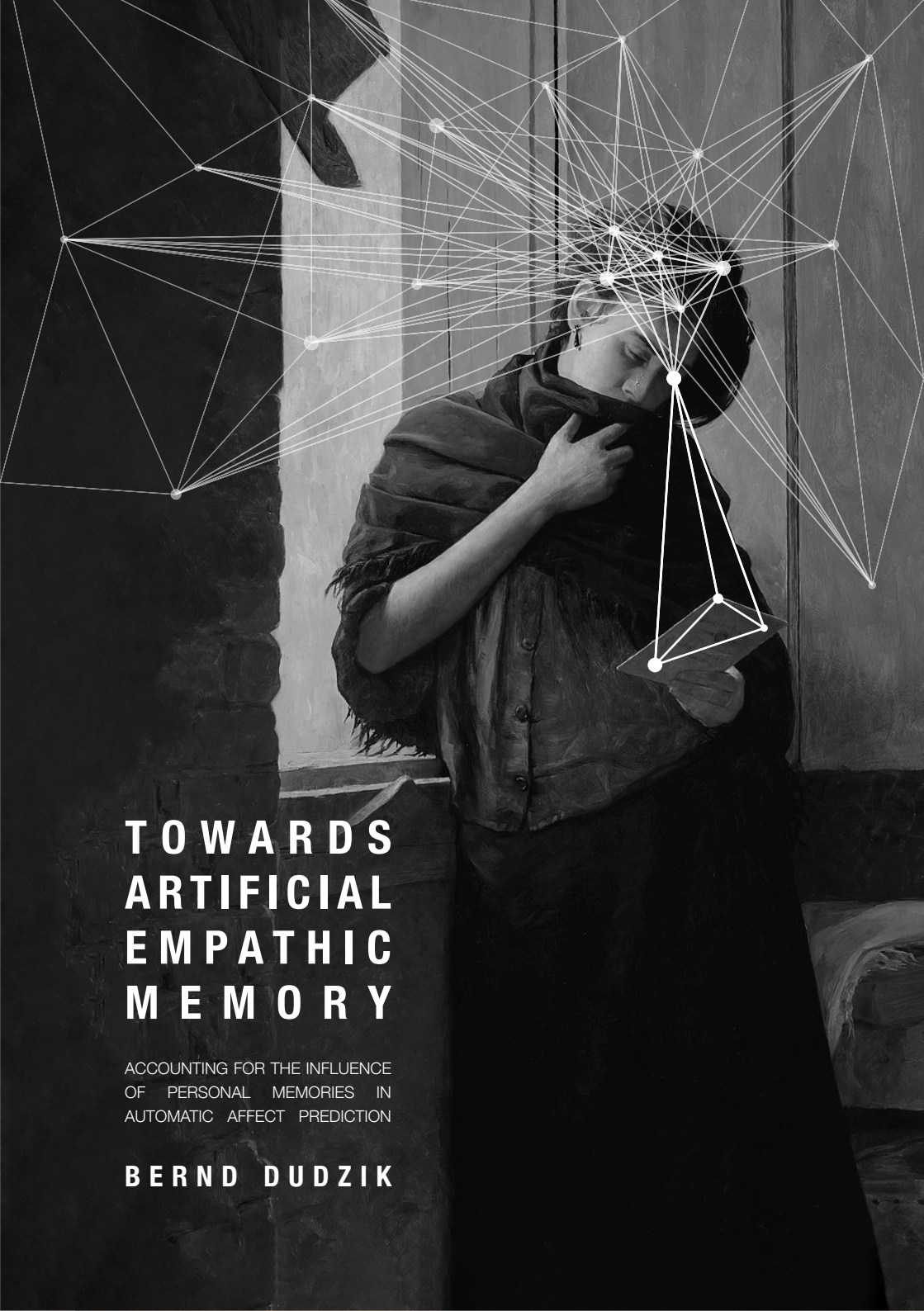
DOI
10.4233/uuid:2d821f6a-67f2-46ed-8518-6e6fa07580d7

Publication date
2021

Document Version
Final published version

Citation (APA)
Dudzik, B. J. W. (2021). *Towards Artificial Empathic Memory: Accounting for the Influence of Personal Memories in Automatic Predictions of Affect*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:2d821f6a-67f2-46ed-8518-6e6fa07580d7

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.



TOWARDS ARTIFICIAL EMPATHIC MEMORY

ACCOUNTING FOR THE INFLUENCE OF PERSONAL MEMORIES IN AUTOMATIC AFFECT PREDICTION

BERND DUDZIK

Towards Artificial Empathic Memory

Accounting for the Influence of Personal Memories in Automatic Affect Prediction

TOWARDS ARTIFICIAL EMPATHIC MEMORY

ACCOUNTING FOR THE INFLUENCE OF PERSONAL MEMORIES IN AUTOMATIC AFFECT PREDICTION

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of Rector Magnificus prof. dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Thursday 4 November 2021, at *15:00* o'clock

door

Bernd Johannes Wilhelm DUDZIK

Master of Science in Mediatechnology, Leiden University, The Netherlands
Born in Singen (Hohentwiel), Germany

This dissertation has been approved by the promoters.

Composition of the Doctoral Committee:

Rector Magnificus,	Chairman
prof. dr. M. A. Neerincx,	Delft University of Technology, Promoter
dr. H. Hung,	Delft University of Technology, Promoter
dr. D. J. Broekens,	Leiden University, Copromoter

Independent Members:

prof. dr. A. Hanjalic,	Delft University of Technology
prof. dr. D.K.J.Heylen,	University of Twente
prof. dr. M. A. Larson,	Radboud University
dr. M. Soleymani,	USC Institute for Creative Technologies, United States
prof. dr. M. J. T. Reinders,	Delft University of Technology, reserve member

This work has been supported by the *4TU research center Humans & Technology (H&T) project (Systems for Smart Social Spaces for Living Well: S4)*.

Keywords: Affect Prediction, Multimodal Modeling, Context, Episodic Memory

Printed by: Druk Tan Heck

Front & Back: Designed by Sara Sallam. Based on *Saudade (1899)*, by Almeida Júnior.

CONTENTS

1

INTRODUCTION

This chapter consists of material adapted from: Dudzik, B., Hung, H., Neerincx, M., & Broekens, J. (2018). *Artificial Empathic Memory*. Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions - EE-USAD'18, 1–8.

1

1.1. EMPATHIC TECHNOLOGY: PREDICTING USER AFFECT TO PERSONALIZE HUMAN-COMPUTER INTERACTIONS

Personalization is about enabling computer systems to autonomously adapt their functionality and behavior based on information about the users currently interacting with them. Its goal is to facilitate interaction experiences that better cater to individuals' specific preferences and needs [1]. For this purpose, such adaptive systems sense and process data about their users as well as their interactions with the system. This data is then used to derive a *User Model* [2], i.e., to identify characteristics of the person in question relevant to the functionality of the application. Based on this model, the system attempts to adapt its behavior adequately.

Traditionally user-models have largely comprised aspects that are expected to remain stable throughout a given interaction, such as a person's general level of experience with a task (e.g., novice vs. expert-users). However, with more widespread use and more complex tasks-environments in which computer systems are being deployed, adaptation requires incorporating more dynamic aspects of interaction and the conditions under which it takes place [1]. In particular, modern computer systems are increasingly expected to make autonomous decisions to support human psycho-social needs. Examples include providing engaging experiences in entertainment [3, 4], and education [5], or supporting mental healthcare [6].

For computer systems to display adaptive or even autonomous behavior in such settings requires the ability to dynamically model and react to users' emotional experience. For this reason, research on *Affective Computing* [7] attempts to equip computers with the ability to predict affective states and responses from available data for purposes of adaptation and personalization. There are different ways in which such information about users' emotions can be valuable for adapting human-computer interactions [8]. First, certain affective states can negatively influence performance in task-critical cognitive processes (such as attention, memory encoding, and decision-making). Computer systems should adequately respond once such circumstances occur or prevent them altogether. For example, an intelligent support system in a car may want to adapt its behavior to angry or agitated drivers to increase road safety for them and others. Secondly, emotional expressions and responses provide implicit feedback about users' appraisal of their current situation [9]. Access to this information can enable computer systems to display adaptive behavior in the future by learning from past interactions, without interrupting the user for explicit evaluations [10]. For example, detecting pleasure or displeasure during exposure to emotional media stimuli, e.g., video material or news articles, can be used by computer systems to learn about users' preferences to provide them with personalized recommendations [11]. Finally, experiencing situations with certain emotional qualities may be a goal in and of itself for people in different circumstances and is crucial for entertainment and emotion regulation [12]. In conclusion, enabling computer systems to predict users' affect is a vital capacity for personalizing human-computer interactions.

In line with these expected benefits, we can group existing technological approaches for automatically predicting affective states from data into two broad strains: (1) *Affect Detection* and (2) *Affective Impact Estimation*. Detecting affect is likely the more prominent endeavor of the two. By and large, approaches for this task exploit audiovisual recordings

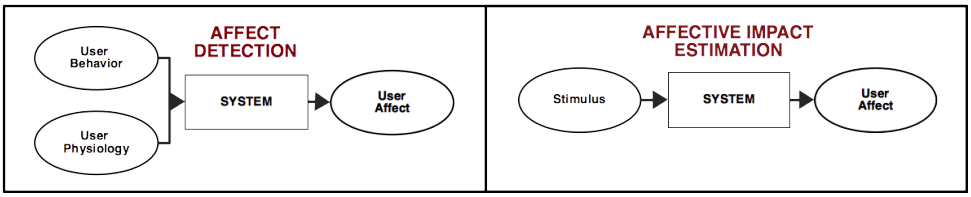


Figure 1.1: Schematic depiction of the two dominant prediction tasks for user affect, together with the input information that they predominantly rely on.

as input data and attempt to identify patterns in behavioral signals (e.g., facial expressions, gestures, or voice activity) associated with certain affective states [13]. However, various internal physiological processes have also been connected by psychological research to emotional responses, such as the activity of the Autonomous Nervous System [14], and current approaches regularly exploit these relationships for automatic affect detection [15]. Estimating affective impact consists of approaches that predict individuals' likely responses to stimuli, potentially before users encounter them. A prime example of this task is *Affective Content Analysis* – which attempts to estimate the emotional impact of exposing users to media by automatically analyzing their content [16].

1.2. CONTEXT AS A CHALLENGE FOR AFFECT PREDICTION

Existing approaches for predicting affect have demonstrated the feasibility of overcoming technological hurdles (e.g., dealing with sensor noise and real-time processing) and display moderate performance in clearly defined research settings [13]. However, for personalizing complex interactions with computers in real-world scenarios, these technologies require both accurate and reliable information about individual users' feelings. Existing approaches for predicting affect struggle to meet this demand due to the high degree of variation in affect elicitation, expression, and perception across different individuals and situations. For example, detection of affect based on only facial expression analysis may provide inaccurate results because the same expression may change its emotional meaning, depending on the context in which a person displays it [17]. Similarly, estimating a specific persons' feelings about a piece of media content may be difficult because it depends on the context under which they engage with it (see, e.g., Soleymani et al.'s findings on mood or time of day [18]).

Context is a notoriously challenging construct to define and has been extensively discussed in research on personalization through adaptive systems (e.g., in a seminal work by Dey et al. [19] focused on ubiquitous computing systems). However, concerning variation in affective processes, we adopt the following working definition proposed by Greenaway et al. [20]:

"*[Context is] a collection of sociocultural forces that shape experience [... These] range from micro-level (intra- and inter-) personal factors that differentiate individuals and groups to macro-level political and historical factors that differentiate cultures.*"

This definition highlights two essential properties of context, namely that (1) it is defined relative to a process or a phenomenon because it shapes it (i.e., a thing that it contextualizes), and (2) that it can be considered and described at different levels of

abstraction (i.e., it follows a hierarchical organization).

The first property highlights its relevance to automatic affect prediction: contextual differences drive variation in affect elicitation, expression, and perception. Consequently, when the context of interactions with a system differs in unanticipated ways, the performance of models for automatic affect predictions may not generalize to this new set of conditions. The second property indicates that the forces constituting context may be similar within particular groupings of individuals (e.g., sharing gender or age) or situations in a nested hierarchy of conditions. Greenaway et al. [20] provide a coarse structure for this hierarchy of context from influences at the micro- to the macro-level: (1) Personal Features, (2) Situational Features, and (3) Cultural Features.

Despite empirical findings from psychology demonstrating a wide variety of influences at each of these levels on human affective processing, existing research on automatic affect prediction has largely neglected them in computational modeling. Instead, it has primarily concentrated on the context-free automatic analysis of a combination of human behavioral signals (see D'Mello and Kory for a review of technical approaches to multimodal affect detection [13]). This relative neglect exists despite researchers within the affective computing community often agreeing on the benefits that context-sensitive predictions would have for the accuracy and robustness of predictions [18, 21–23].

Two potential reasons for hindering progress are that it is both (1) unclear what constitutes contextual influences that are *effective* for improving automated prediction technology, as well as (2) how to *feasibly* obtain and incorporate relevant information about them in a technological system (see also Hammal and Suarez for a discussion of challenges for context-sensitive affect prediction [21]) .

The premise of the research presented in this dissertation is that overcoming these specific hurdles requires a concentrated research effort involving the interdisciplinary collaboration between social and computer scientists. In particular, we are convinced that progress is dependent on the systematic exploration of those contextual influences within computational modeling activities that have been previously identified as relevant for human affective processes in empirical work from the social sciences. This dissertation is contributing to such an exploration. *It investigates the effectiveness and potential feasibility of accounting for one particularly important context in affect predictions: the recollection of personal memories.*

1.3. THE IMPORTANCE OF PERSONAL MEMORIES FOR AFFECT PREDICTION

An essential part of being an individual is our personal history, in particular, our personal memories. They form an essential contextual driver for our emotional and cognitive interpretation of what is *currently happening*, including interactions with computers. However, current approaches for personalizing interactions with computers are neither aware of what memories are triggered in users nor their emotional interpretations of those memories.

This section first presents evidence for the crucial role of personal memories in human affective processing. We then describe examples of how unawareness of their influence may prevent computer systems from correctly predicting user affect.

1.3.1. PERSONAL MEMORIES AS A DRIVER OF AFFECT

Moments in which human beings re-experience specific events from within their personal history are known as *Episodic Memories* [24]. These recollections typically include a sense of the time and the place at which remembered events have occurred, as well as potentially vivid visual imagery [24]. People intentionally engage in recollecting such memories for various functions, such as planning and decision-making, and as a resource to fulfill deep-rooted psycho-social needs [25, 26]. For example, revisiting shared experiences is a crucial component of maintaining intimacy with loved ones, while disclosing anecdotes from one's past can be a vital mechanism for forging new bonds with others. However, apart from being voluntarily brought to mind, personal memories may also be triggered involuntarily by external triggers in our current environment [27]. Such involuntary recollections frequently occur in everyday life [28].

Empirical evidence indicates that memories brought to mind either voluntarily or involuntarily may contain strong affective associations that influence our present emotional interpretations of situations [29]. Personal memories' potential impact is further underlined by their recollection being used for emotion induction in empirical research (see, e.g., Mills and D'Mello [30]). Moreover, memories that are triggered involuntarily seem to result in a more immediate and intensive emotional impact than those that are intentionally recalled, likely because they arrive suddenly and without individuals' having time to prepare [28]. As such, especially personal memories triggered in this fashion can form a highly dynamic contextual influence on individuals' affective state, forming spontaneously occurring mental stimulus events with a substantial degree of impact.

Interactions with media content may be particularly affected by memory-influences. For example, empirical research has identified that musical pieces can act as potent triggers for the recollection of events from a listener's past [31, 32]. Moreover, the emotional tone associated with the memories elicited influences how listening to a piece feels [33, 34].

1.3.2. RECOLLECTION-UNAWARENESS: LIMITATIONS RESULTING FROM DISCOUNTING PERSONAL MEMORIES IN AFFECT PREDICTION

Despite psychological research underlining the importance of personal memories for shaping affective experience, existing automatic affect predictions have ignored it thus far. However, without accounting for this highly dynamic influence, these are likely to fail at providing accurate and useful estimates of users' affect in many different application domains. The reason for this is that they are unable to predict *if* memories impact a user's' feelings in a given situation at all, *what* memories impact them, and *how* these memories impact them. We claim that being *recollection-unaware* in such a fashion severely limits the degree to which computers can display meaningful empathic understanding or behavior towards their users since they cannot relate to the influence these internal stimuli exercise during interactions. In the following, we discuss a series of examples from different application domains, highlighting the consequences for computer systems that want to anticipate the affective impact of their actions on users or rely on detected affect as a feedback signal to adjust their behavior in future human-computer interactions.

Reminiscence Support Systems: Researchers have displayed an ongoing interest in designing technologies to support and encourage reminiscing activities through the provision of personally relevant media content [35–39]. Such activities have been shown to fulfill numerous psycho-social functions for individuals engaging in it [26] and improve their subjective well-being [40]. However, because existing approaches to these systems are recollection-unaware, they are very limited in their capacity to trigger personal memories aligned with their specified functionality or their users' desires. Instead, to function reliably, they need to be capable of providing highly personalized experiences, i.e., they require the recommendation of multimedia stimuli that are emotionally meaningful for a specific individual user, in light of his or her past as accessible and experienced at the time of interaction. Technologies cannot achieve this without explicitly accounting for the influence of the triggered personal memories when predicting a stimulus's affective impact on their users.

Recommender Systems: Beyond recommending stimuli for reminiscing or reminding, interactions with any form of recommender system can benefit from recollection-awareness. These benefits relate both to what suggestions these systems make and the means that they choose for presenting their recommendations. Contemporary versions of such systems already integrate some contextual features to make intelligent recommendations [41], for example, location and time (see, e.g., Saiph Savage et al. [42]). As such, they can, for example, suggest dinner locations with the understanding that it is lunchtime and where the nearest dining locations are. However, these recommendations cannot consider the local haunts of an ex-lover or that a restaurant is important, because a meaningful family celebration for that individual has occurred there. In this case, how a user is experiencing the recommendation is highly influenced by the personal memories associated with a particular dinner location. This association may change over time due to other factors than the person's experience at those locations. Moreover, the form in which a system provides a recommendation (e.g., involving a photo) may accidentally trigger personal memories that influence its experience significantly. Consequently, without understanding the potential influence of personal memories, it may be practically impossible to anticipate a particular recommendation's emotional impact on an individual.

Social Robotics: Since people often bond about reflecting, reliving, and sharing past emotional experiences (see, e.g., [43]), it seems important that social robots can strategically refer to events from their shared past with a user to do the same [44]. However, without the capacity to estimate the emotional significance of a past event (do they even care?) and its impact on a user's current affective state (does it result in a pleasant experience?), these systems will not be able to mimic this human capacity. Similarly, without an understanding of when a situation triggers memories in humans and their emotional impact on them, social robots will have to solely rely on directly observable cues to interpret their affective state (e.g., by analyzing facial expressions). Moreover, these systems will be unable to reason about potential causes for changes in affective state meaningfully and take these into account in their actions (e.g., why did a visit to this specific place cause sadness, but a visit to similar other places did not?).

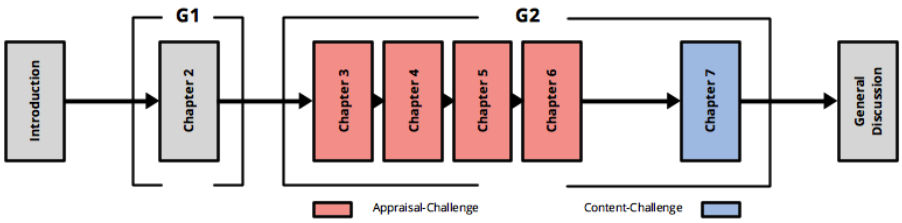


Figure 1.2: Schematic depiction of the research objectives addressed in the different chapters of this dissertation (G1 and G2). Red and blue colored chapters relate to specific prediction-challenges for recollection-aware modeling of user affect.

The examples discussed so far are just a small selection of potentially impacted computer systems limited by a lack of consideration for the impact of personal memories. Because recollections of our past form such an essential part of human cognitive-affective functioning, it is easy to envision numerous other scenarios where users' experiences of interactions could be improved by providing computer systems with some degree of awareness for them.

1.4. THESIS OBJECTIVES AND CONTRIBUTIONS

Motivated by the potential benefits for personalization of a broad range of human-computer interactions, the primary objectives of this dissertation are

G1 the identification of the information that is necessary for a computer system to facilitate recollection-aware modeling of user affect, as well as the additional prediction challenges that need to be solved for providing this information, and

G2 the evaluation of the *effectiveness* and *feasibility* of addressing these prediction challenges in particular application domains.

A schematic overview of how the different chapters of the thesis relate to these objectives can be seen in *Figure 1.2*. In the following, we outline this dissertation's contributions in pursuit of each of these goals and their underlying research questions.[1]

A COMPUTATIONAL ARCHITECTURE AND RESEARCH FRAMEWORK

In *Chapter 2*, we provide a systematic decomposition of the task of recollection-aware affect prediction into a series of specific prediction challenges. We argue that recollection-aware affect prediction requires information about (1) When memories are triggered in a user (REceptiveness-Challenge), (2) what their content is (Content-Challenge), and (3) how interpreting this content is influencing the users' current affective experience of

[1] Please note that the content chapters of this dissertation (i.e., Chapters 2-7) are based on existing publications. We have chosen to leave these chapters largely in the original form in which they were created. While this means that there are certain overlaps and repetitions in them (e.g., introduction, related works, dataset descriptions), this means that they can be read mostly independently from each other and in any desired order.

the memory content, and the situation in which it is recollected (APpraisal-Challenge). We refer to these collectively as the RECAP problem and propose an *Artificial Empathic Memory (AEM)* to address it. The AEM is a computational architecture for the simulation of viewers' cognitive-affective memory processing with a dedicated component for addressing each aspect of the RECAP problem (i.e., by providing the required information through solving a relevant prediction task). The chapter outlines the psychological foundations for the architecture components, describes challenges for instantiating them in computational systems, and existing technological approaches to form the starting points for overcoming these. Overall, the research presented in this chapter is guided by the following questions:

> **RQ1a:** What prediction challenges need to be addressed to provide computer systems with the necessary information for capturing the influence of personal memories when predicting user affect?

> **RQ1b:** How can these challenges be approached using existing knowledge from the social sciences and technological research?

In addition, the AEM architecture also serves as a conceptual framework for structuring research efforts on recollection-aware affect prediction, including the work presented in this dissertation. Concretely, *Chapters 3 to 6* explore the appraisal-challenge for predicting users' responses to video content. In contrast, *Chapter 7* describes research addressing the content-challenge with data collected via ubiquitous sensing (See *Figure 1.2* for a schematic representation).

EXPLORATION OF THE APPRAISAL-CHALLENGE IN AUTOMATIC PREDICTIONS OF VIDEO-INDUCED EMOTIONS

For exploring the appraisal challenge, we focus on predicting viewers' emotional responses to video content. We opted for this setting because this task is one of the primary research areas for predicting user affect (e.g., [16]). Moreover, such predictions are likely to be relevant for real-world applications in the foreseeable future (e.g., for content recommendations [45]). Finally, predicting affect in this setting provides a relatively constrained scenario for exploring context-sensitive predictions of user affect, compared to the complex structure surrounding social interactions (see, e.g., the review of potential influences in such a setting described in Dudzik et al. [46]).

A Multimodal Dataset for Modeling Affect and Memory Processes: We collected a multimodal dataset of individuals' cognitive-affective responses to video stimuli that we use for all empirical investigations and modeling activities related to the appraisal-challenge. We refer to this corpus as the *Mementos* dataset and provide a detailed description of its contents in *Chapter 3*. To the best of our knowledge, it is the first dataset developed for computational modeling of interactions between memory processing and affective interpretation of video-stimuli. Particular challenges for its collection were related to

capturing the influence of personal memories on video-induced emotions in an ecologically valid setting while also obtaining data that facilitates computational modeling. The guiding research question for the dataset construction was:

> **RQ2:** How should a protocol for data collection be designed to capture the influence of personal memories in an ecologically valid setting while also obtaining data that facilitates computational modeling?

Investigating the Impact of Personal Memories: We opted for modeling responses to video stimuli, instead of other types of media – such as photographs or audio recordings –, because videos have a duration over time and are capable of carrying complex multimodal content. This fidelity makes videos a more ecologically valid choice than other media since they arguably represent human stimulus experiences in everyday life more closely. Moreover, this richness also might enable videos to spark associations with memories in different ways, e.g., connections to background music or visual similarity. However, despite psychology having identified personal memories as an essential driver for emotional responses to media-content (see, e.g., Scherer's discussion on the topic of music [47]), specific empirical insights about effects in a video-viewing setting are currently not available. Hence, one of our key research questions is to explore the influence of recollections on emotional responses in this setting:

> **RQ3:** What influences do the occurrence and emotional meaning of personal memories have on video-induced emotions?

Using the Mementos dataset, *Chapter 4* presents a series of statistical analyses that explore and quantify how the occurrence and emotional appraisal of personal memories influence viewers' emotional responses. The upshot of these investigations is to understand how far the general theoretical motivation for investigating personal memories is substantiated by empirical evidence in this particular setting. If there is no substantial effect of personal memories on video-induced emotions, then accounting for them in automatic predictions is ineffective.

Exploration of Effectiveness and Feasibility for Automatic Predictions: Because recollected memory content cannot be directly observed with technological sensors, addressing the appraisal-challenge in automatic predictions poses a substantial challenge. As such, our research's principal focus is exploring the *effectiveness* of modeling memory appraisals for improving predictions of user affect and establishing the *feasibility* of doing so through automatic analysis of data that is potentially available to an automatic system.

A crucial step in assessing the effectiveness of addressing the appraisal-challenge for predicting affect, is to compare it to existing alternatives that provide context for predictions. One such alternative is a static profile with potentially relevant information about users, such as demographic characteristics or personality. Compared to providing computer systems with information about memory content triggered in viewers, such a profile is comparatively easy to assemble using explicit self-reports. Consequently, should

it offer similar performance benefits for predicting video-induced emotions, then accounting for memory influences in automatic predictions holds no additional advantages and can be considered an ineffective enterprise. We assess this as part of the statistical analyses described in *Chapter 5*, guided by the following research question:

> **RQ4:** Do personal memories provide information about emotional responses to videos that go beyond those offered by relevant static user characteristics (i.e., demographics, personality, or mood)?

In *Chapter 6*, we then explore the feasibility of addressing the appraisal-challenge by analyzing user-data that might be available to a technological system. In particular, we attempt to model the influence of personal memories on video-induced emotions based on automatic analysis of free-text descriptions of memory content that is part of the Mementos dataset. All of our technical explorations can be considered to occur in an Affect Detection-setting, rather than Affective Impact Estimation. This is the case since viewers provided the analyzed memory descriptions as part of their response to the video stimuli, not before. Consequently, we are guided by the following research question:

> **RQ5:** Can we detect video-induced emotions based on an automatic analysis of free-text descriptions of their content?

Building on this automatic analysis of memory content, we present a series of machine learning experiments that explore its benefits for predicting video-induced emotions when combined with existing paradigms for this task that do not consider additional context. Concretely, we explore the benefits of access to personal memories for: (1) Video-Affective Content Analysis (VACA) – where typically only the eliciting videos' audiovisual context is considered –, and (2) Facial Behavior Analysis – where typically only changes in a viewers' facial expressions are considered. Here our research questions are:

> **RQ6a:** Does automatic analysis of memory descriptions improve the performance of detections compared to those based on context-free analysis of video content?

> **RQ6b:** Does automatic analysis of memory descriptions improve the performance of detections compared to those based on context-free facial behavior analysis?

EXPLORATION OF THE CONTENT CHALLENGE BASED ON LIFELOG DATA

Our explorations of the appraisal-challenge have relied on explicit descriptions of recollected memory content for automatic analysis. However, a crucial part of achieving recollection-aware affect predictions is the broad availability of information about individuals' past that forms the potential content of recollected memories (i.e., a resource for addressing the content-challenge). Lifelogging is a form of ubiquitous data collection

that holds the potential to provide computational models with such a resource. It aims to collect, aggregate, and organize personal data into a comprehensive personal timeline for access and retrieval [48]. For example, researchers have explored the usage of small cameras for continuous picture-taking from a first-person perspective for this purpose (see Bolaños et al. for a recent technical overview [49]). In *Chapter 7*, we move on towards a case-study touching on the feasibility of lifelog data to address the content-challenge. For this purpose, we model individuals' recollection processes based on lifelog-data about their social interactions in the past and prompts that they were provided with for recall at a later point in time. Concretely, we explore identifying segments in a stream of lifelog-data that correspond to memory content that individuals have recollected. As such, our final research question is:

> **RQ7:** Can we identify segments in the lifelog data of individuals that correspond to the memory content they have recollected?

1.5. LIST OF PUBLICATIONS RELATED TO THE THESIS

- **Dudzik, B.**, Neerincx, M., Hung, H., & Broekens, J. (2018). *Artificial empathic memory: Enabling media technologies to better understand subjective user experience.* EE-USAD 2018 - Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions, Co-Located with MM 2018. https://doi.org/10.1145/3267799.3267801

- **Dudzik, B.**, Hung, H., Neerincx, M. A., & Broekens, J. (2021). *Collecting Mementos: A Multimodal Dataset for Context-Sensitive Modeling of Affect and Memory Processing in Responses to Videos.* IEEE Transactions on Affective Computing, 3045(c), 1–1. https://doi.org/10.1109/TAFFC.2021.3089584

- **Dudzik, B.**, Hung, H., Neerincx, M., & Broekens, J. (2020). *Investigating the Influence of Personal Memories on Video-Induced Emotions.* Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 53–61. https://doi.org/10.1145/3340631.3394842

- **Dudzik, B.**, Broekens, J., Neerincx, M., & Hung, H. (2020). A Blast From the Past: Personalizing Predictions of Video-Induced Emotions using Personal Memories as Context. ArXiv. https://arxiv.org/abs//2008.12096

- **Dudzik, B.**, Broekens, J., Neerincx, M., & Hung, H. (2020). *Exploring Personal Memories and Video Content as Context for Facial Behavior in Predictions of Video-Induced Emotions.* Proceedings of the 2020 International Conference on Multimodal Interaction, 10(20), 153–162. https://doi.org/10.1145/3382507.3418814

- **Dudzik, B.**, Olenick, J., Broekens, J., Chang, C. H., Hung, H., Neerincx, M., & Kozlowski, S. W. J. (2018). *Discovering digital representations for remembered episodes from lifelog data.* Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, MCPMD 2018. https://doi.org/10.1145/3279810.3279850

1

REFERENCES

[1] G. Fischer, *Context-aware systems,* in *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12*, AVI '12 (ACM, New York, NY, USA, 2012) p. 287.

[2] A. Kobsa, *Generic User Modeling Systems,* in *The Adaptive Web*, edited by N. W. e. In: Brusilovsky P., Kobsa A. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007) pp. 136–154.

[3] A. Hanjalic, *Extracting Moods from Pictures and Sounds: Towards truly personalized TV,* IEEE Signal Processing Magazine **23**, 90 (2006).

[4] M. Soleymani and M. Pantic, *Emotionally Aware TV,* Proceedings of TVUX-2013: Workshop on Exploring and Enhancing the User Experience for TV at ACM CHI 2013 (2013).

[5] A. C. Graesser, M. W. Conley, and A. Olney, *Intelligent tutoring systems.* in *APA educational psychology handbook, Vol 3: Application to learning and teaching.* (American Psychological Association, Washington, 2012) pp. 451–473.

[6] F. Burger, M. A. Neerincx, and W.-P. Brinkman, *Technological State of the Art of Electronic Mental Health Interventions for Major Depressive Disorder: Systematic Literature Review,* Journal of Medical Internet Research **22**, e12599 (2020).

[7] R. Picard, *Affective Computing for HCI.* HCI (1) (1999).

[8] E. Hudlicka, *To feel or not to feel: The role of affect in human-computer interaction,* International Journal of Human Computer Studies **59**, 1 (2003).

[9] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, *Appraisal theories of emotion: State of the art and future development,* Emotion Review **5**, 119 (2013).

[10] J. Broekens, *Emotion and Reinforcement: Affective Facial Expressions Facilitate Robot Learning,* in *Artifical Intelligence for Human Computing*, Vol. 4451 LNAI (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007) pp. 113–132.

[11] M. Soleymani and M. Pantic, *Human-centered implicit tagging: Overview and perspectives,* in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* (IEEE, 2012) pp. 3304–3309.

[12] A. Bartsch, *Emotional Gratification in Entertainment Experience. Why Viewers of Movies and Television Series Find it Rewarding to Experience Emotions,* Media Psychology **15**, 267 (2012).

[13] S. K. D'mello and J. Kory, *A Review and Meta-Analysis of Multimodal Affect Detection Systems,* ACM Computing Surveys **47**, 1 (2015).

[14] S. D. Kreibig, *Autonomic nervous system activity in emotion: a review.* Biological psychology **84**, 394 (2010).

[15] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, *Physiological signals based human emotion recognition: A review,* in *Proceedings - 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications, CSPA 2011* (IEEE, 2011) pp. 410–415.

[16] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, *Affective Video Content Analysis: A Multidisciplinary Insight,* IEEE Transactions on Affective Computing **9**, 396 (2018).

[17] 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,* Psychological Science in the Public Interest **20**, 1 (2019).

[18] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, *Corpus Development for Affective Video Indexing,* IEEE Transactions on Multimedia **16**, 1075 (2014), arXiv:1211.5492 .

[19] A. K. Dey, *Understanding and Using Context,* Personal and Ubiquitous Computing **5**, 4 (2001).

[20] K. H. Greenaway, E. K. Kalokerinos, and L. A. Williams, *Context is Everything (in Emotion Research),* Social and Personality Psychology Compass **12**, e12393 (2018).

[21] Z. Hammal and M. T. Suarez, *Towards context based affective computing introduction to the third international CBAR 2015 workshop,* in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (IEEE, 2015) pp. 1–2.

[22] S. Wang and Q. Ji, *Video Affective Content Analysis: A Survey of State-of-the-Art Methods,* IEEE Transactions on Affective Computing **6**, 410 (2015).

[23] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, *Affective Computing for Large-scale Heterogeneous Multimedia Data,* ACM Transactions on Multimedia Computing, Communications, and Applications **15**, 1 (2020), arXiv:1911.05609 .

[24] M. A. Conway, *Episodic memories,* Neuropsychologia **47**, 2305 (2009).

[25] S. Bluck, N. Alea, and B. Demiray, *You Get What You Need,* in *The Act of Remembering* (Wiley-Blackwell, Oxford, UK, 2010) pp. 284–307.

[26] J. Dean Webster, *The reminiscence circumplex and autobiographical memory functions,* Memory **11**, 203 (2003).

[27] D. Berntsen, S. R. Staugaard, and L. M. T. Sørensen, *Why am I remembering this now? Predicting the occurrence of involuntary (Spontaneous) episodic memories,* Journal of Experimental Psychology: General **142**, 426 (2013).

[28] D. Berntsen, *The Unbidden Past,* Current Directions in Psychological Science **19**, 138 (2010).

[29] H. Baumgartner, M. Sujan, and J. R. Bettman, *Autobiographical Memories, Affect, and Consumer Information Processing,* Journal of Consumer Psychology **1**, 53 (1992).

[30] C. Mills and S. D'Mello, *On the Validity of the Autobiographical Emotional Memory Task for Emotion Induction,* PLoS ONE **9**, e95837 (2014).

[31] M. D. Schulkind, L. K. Hennis, and D. C. Rubin, *Music, emotion, and autobiographical memory: They're playing your song,* Memory & Cognition **27**, 948 (1999), arXiv:arXiv:1011.1669v3 .

[32] L. Jäncke, *Music, memory and emotion,* Journal of Biology **7**, 21 (2008).

[33] M. Zentner, D. Grandjean, and K. R. Scherer, *Emotions evoked by the sound of music: Characterization, classification, and measurement.* Emotion **8**, 494 (2008).

[34] H. Baumgartner, *Remembrance of things past: Music, autobiographical memory, and emotion,* in *NA - Advances in Consumer Research*, Vol. 19, edited by J. F. Sherry Jr. and B. Sternthal (1992) pp. 613–620.

[35] E. van den Hoven, *A future-proof past: Designing for remembering experiences,* Memory Studies **7**, 370 (2014).

[36] E. van den Hoven and B. Eggen, *The Cue Is Key,* Zeitschrift für Psychologie **222**, 110 (2014).

[37] D. Cosley, V. Schwanda, S. T. Peesapati, J. Schultz, and J. Baxter, *Experiences with a publicly deployed tool for reminiscing,* in *CEUR Workshop Proceedings*, Vol. 499 (2009) pp. 31–36.

[38] K. O'Hara, J. Helmes, A. Sellen, R. Harper, M. Ten Bhömer, and E. Van Den Hoven, *Food for talk: Phototalk in the context of sharing a meal,* Human-Computer Interaction **27**, 124 (2012).

[39] V.-T. Nguyen, K.-D. Le, M.-T. Tran, and M. Fjeld, *NowAndThen,* in *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia - MUM '16*, MUM '16 (ACM, New York, NY, USA, 2016) pp. 159–168.

[40] F. B. Bryant, C. M. Smart, and S. P. King, *Using the Past to Enhance the Present: Boosting Happiness Through Positive Reminiscence,* Journal of Happiness Studies **6**, 227 (2005).

[41] G. Adomavicius and A. Tuzhilin, *Context-Aware Recommender Systems,* in *Recommender Systems Handbook* (Springer US, Boston, MA, 2015) pp. 191–226, arXiv:arXiv:1011.1669v3 .

[42] N. Saiph Savage, M. Baranski, N. Elva Chavez, and T. Höllerer, *I'm feeling LoCo: A Location Based Context Aware Recommendation System,* in *Lecture Notes in Geoinformation and Cartography*, 199599 (Springer Berlin Heidelberg, 2012) pp. 37–54.

[43] N. Alea and S. Bluck, *Why are you telling me that? A conceptual model of the social function of autobiographical memory,* Memory **11**, 165 (2003).

[44] M. Y. Lim, *Memory Models for Intelligent Social Companions,* in *Studies in Computational Intelligence*, Vol. 396 (2012) pp. 241–262.

[45] M. Soleymani and M. Pantic, *Emotionally Aware TV,* in *Proceedings of TVUX-2013: Workshop on Exploring and Enhancing the User Experience for TV at ACM CHI 2013*, April 2013 (2013).

[46] B. Dudzik, M.-P. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong, *Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases,* in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 206–212.

[47] K. R. Scherer, *Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms? And How Can We Measure Them?* Journal of New Music Research **33**, 239 (2004).

[48] C. Gurrin, A. F. Smeaton, and A. R. Doherty, *LifeLogging: Personal Big Data,* Foundations and Trends® in Information Retrieval **8**, 1 (2014).

[49] M. Bolanos, M. Dimiccoli, and P. Radeva, *Toward Storytelling From Visual Lifelogging: An Overview,* IEEE Transactions on Human-Machine Systems , 1 (2017).

2

An Architecture for Artificial Empathic Memory: Enabling Computers to Better Understand Subjective User Experience

This chapter is based on: Dudzik, B., Neerincx, M., Hung, H., & Broekens, J. (2018). *Artificial empathic memory: Enabling media technologies to better understand subjective user experience.* EE-USAD 2018 - Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions, Co-Located with MM 2018. https://doi.org/10.1145/3267799.3267801

ABSTRACT

While personal memories are a crucial driver for individuals' emotional experience, their influence is currently not considered in automatic affect predictions, i.e. these approaches are recollection-unaware. In this chapter, we identify the information about individuals' recollections that needs to be accessible by automatic systems in order to account for it in automatic predictions. Providing this information forms a set of substantial prediction challenges. As a way for computer systems to overcome these challenges for predicting affect, we propose the development of an Artificial Empathic Memory (AEM) of their users. We describe a psychologically inspired architecture, examine the challenges to be solved, and highlight how existing research can become a starting point for overcoming them.

2.1. THE RECAP PROBLEM

In this section, we argue that the influence of personal memories on the affective experience of interactions with computers pose three fundamental prediction challenges that need to be addressed by technologies to provide recollection-aware modeling of user affect:

- Knowing the conditions that trigger episodic memories. We refer to this as the *challenge of receptiveness*, i.e., whether an individual is receptive to remembering episodes in the current situation;
- Knowing what episodic memories of the user are triggered by the current situation. We refer to this as the *challenge of content*, i.e., which memories are triggered by the current situation;
- Knowing the emotional interpretation of a remembered event and its impact on both the interpretation of the current situation and emotional state of the user. We refer to this as the *challenge of appraisal*, i.e., how does the current situation, including the triggered episodic memories, feel for the individual.

We summarize these challenges as the *RECAP problem (REceptiveness, Content, and APpraisal)*.

Approaches for user modeling in human-computer interactions that fail to address the RECAP problem are unable to provide reliable estimations of a user's affective experiences for personalization. This is because they are unable to predict *if* memories impact the user's experience of the current situation at all, *what* memories impact it, and *how* they impact it. In a very real sense, being *recollection-unaware* in such a fashion makes technologies lack empathy, since they cannot relate in any way to the influence that having access to a personal history exercises on their users' feelings.

2.2. AN ARCHITECTURE FOR ARTIFICIAL EMPATHIC MEMORY

Human beings have an innate ability to estimate how other people think and feel in response to events in their environment [1]. An important part of this empathic understanding is the cognitive-affective reasoning by which a person simulates the mental states of others, based on prior (shared-) experiences and general knowledge [2]. This is

often referred to as *Theory of Mind*. Empirical findings suggest that the more familiar one person is with another, the more likely they are to gain accurate insights into how that other person feels [3]. In essence, to achieve true personalization of their interactions, applications need to possess a rudimentary theory of mind of their users. This would enable them to simulate what a user is actually thinking and feeling. Building an accurate and usable artificial theory of mind is, however, a bridge too far in the context of reliable computational modeling.

In this chapter we argue that an important subset of that can already help to address the RECAP problem for providing personalized experiences. To that end, we propose a computational architecture for an *Artificial Empathic Memory (AEM)*. It provides applications with the ability to predict the user's experiences of a situation (including the system's actions) while taking into account the episodic memories that are so important for forming his/her personal interpretation of it.

We argue, that for each of the three challenges comprising the RECAP problem there is a suitable psychological theory that can form the foundation of a functional component to address it computationally. In the following, we provide an overview over each of these components. In particular, we outline how they interact with each other to detect the individual's *attentional engagement* in a present activity, predict the *associative strength* existing between external stimuli and episodic memories, and finally predict their impact on the *emotional experience* of individuals. See also Fig. 2.1 for an overview.

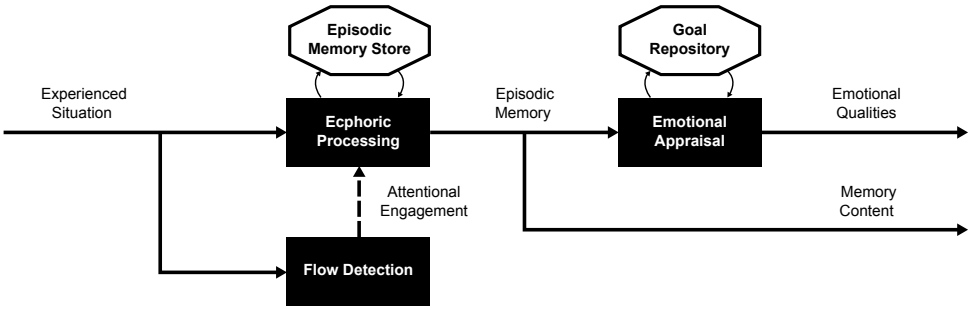


Figure 2.1: An overview of the functional components of the AEM Architecture

- **Flow Detection Component:** The input of this module consists of features describing a user's current activity and state, while the output is the degree of *attentional engagement* that is experienced. This value modifies the operation of the ecphoric processing-module: a low degree of attentional engagement results in a low activation threshold for episodic memories, biasing the Ecphoric Processing Component to propose candidates for recollection.
- **Ecphoric Processing Component:**. The input of this component is the current situation (state + activity). It extracts (a subset of) the user's current situation as an *Episode* in a representation that allows associative strength to be calculated (e.g. a vector of features). Then it determines the *associative strength* between that encoding of the situation and all available episodes in the *Episodic Memory Store*.

The outcome of this operation are one or more *Episodic Memories*. The *Episodic Memory Store* forms an important resource for this process. It is a database that contains a collection of information about personal events from a user's past in the form of encoded situations that we refer to as *Episodes*.

- **Emotional Appraisal Component:** This module simulates a series of cognitive-affective processes that determine the emotional quality of experiencing an episodic memory. It takes a representation of an *Episodic Memory* as input and outputs a representation of the *Emotional Qualities* of its experience. An important resource needed for this component is a *Goal Repository*, containing information about existing concerns and motivations of the user who is being modeled.

In the remainder of this chapter, we provide an outline of psychological theories that we have chosen to form the foundations for these functional components. Furthermore, we highlight existing computational work in line with this psychological basis. As such, our argument is that an AEM is not only needed for addressing the RECAP problem, as explained above, but also feasible in the future, given sufficient efforts. Finally, we discuss conceptional and technological challenges that need to be tackled to instantiate each functional component of the AEM.

2.3. FLOW DETECTION AND THE CHALLENGE OF PREDICTING RECEPTIVENESS

2.3.1. PSYCHOLOGICAL BACKGROUND

Findings of empirical studies investigating the emergence of episodic memories in everyday life have demonstrated that they have a tendency to occur in situations where a person's attention is not fully immersed in an ongoing activity [4, 5]. In this section, we argue that *Flow* is a useful psychological concept to understand and model the degree to which a user's current situation gives rise to such attentional engagement.

The concept of flow describes a state of mind in which a person is so absorbed in performing an activity that there is no room for other thoughts to emerge [6]. A requirement for flow experiences to emerge, is that an ongoing situation holds a balance between the challenges that it presents to a person and his or her perceived ability to cope with them [7]. Importantly, situations that are experienced as lacking in challenge result in states of cognitive under-stimulation, e.g. boredom [8]. Here, individuals' attentional resources are no-longer fully invested in the activities they are undertaking, thus creating conditions that are more favorable for episodic memories to emerge.

In summary, Flow is a concept that is widely used and empirically well established across a broad variety of disciplines. It provides a suitable theoretical framework to characterize how individuals' degree of attentional engagement varies under specific circumstances, which in turn modulates individuals' tendency for episodic recollection.

2.3.2. COMPUTATIONAL APPROACHES TOWARDS FLOW DETECTION

A large body of work on flow and engagement detection exists, within the domains of entertainment and education computing. For example, research on detecting tutoring engagement showed initial successes at discriminating between flow-relevant states of

boredom, frustration and confusion in learners [9]. In the adaptive gaming domain, automatic detection of boredom and frustration was also shown to be feasible [10, 11]. In some cases, these attempts reached a reported accuracy of over 90 percent in post-hoc classification of engagement and frustration based on recorded visual and game-play features [11]. Further, in the field of Human Robot Interaction, initial investigations have shown the possibility to detect engagement based on task and interaction-related features [12], in essence replicating findings in the gaming and e-learning/tutoring domain. Finally, research in the field of interruptibility detection (e.g. [13]) strongly relates to detecting flow and engagement based on a user's context as captured by ubiquitous sensing technology [14]. In essence, these different areas seem to converge on similar ideas, namely, that it is both important and computationally feasible to detect task engagement in users. With the right focus, we believe these techniques can be extended to engagement measures that are correlated with the emergence and intensity of episodic recollections.

2.3.3. CHALLENGES IN FLOW DETECTION

Important research challenges remain. First, significant sensing abilities are needed to detect engagement in users, in particular when focusing on social signals. Pupil-dilation might be an interesting alley for future research, as it seems to correlate with, for example, high temporal resolution attention dynamics [15]. As such, it might be an easy-to-detect, uni-modal option for the detection of flow. As eye-tracking can now be reliably done using machine learning on data coming from standard cameras embedded in mobile devices [16], it is to be expected that pupil dilation detection too becomes feasible in the near future. This opens up the possibility to detect attentional engagement in real time on standard customer devices.

Another challenge hinted at by flow theory is that activities may result in varying degrees of attentional engagement for different individuals, because these do not experience the same degree of challenge. Detection in these circumstances can probably be enhanced with personalized engagement models. Information about a user's specific skills or interest in particular activities might help a computational model of flow-processing to become more accurate in detecting momentary engagement.

Finally, flow detection will also need to be taken into account when compiling data traces into digital records describing persons' episodic stores. This is because the information sensed by technological monitoring may not be aligned with the deployment of attentional resources by a person in the same situation. It may therefor provide a description of the events in question that is strikingly different from that person's memories. Research on modeling human attentional focus (e.g. [17]) holds potential for improvement of this circumstance, e.g. by enabling applications to construct a model of a situation that corresponds more closely to the user's perception of it.

2.4. ECPHORIC PROCESSING AND THE CHALLENGE OF PREDICTING EPISODIC CONTENT

2.4.1. PSYCHOLOGICAL BACKGROUND

The degree of association between a present situation and an instance in a person's past plays an essential role in the emergence of episodic memories with a specific content in

contemporary psychological models of human memory (e.g. [18, 19]). For example, it is understood that the potential of a present situation to cause an episodic memory of a specific past event (i.e. to act as a *memory cue* for it), is dependent on its similarity to the context under which that event was originally committed to memory [20, 21]. The greater this overlap, the more likely it is to come to mind. Consequently, the associative strength of external stimuli both influences whether something comes to mind, as well what something is. However, the nature of the associations linking external stimuli to past events can take numerous forms and exist at different levels of abstraction. They can range from purely perceptual similarities between cues and elements of a past episode to associations that exist solely at a conceptual level [22].

One way to conceptualize the process of how stimuli act as cues for recollections of specific events has been proposed by Tulving [23]: in an initial phase called *ecphory*, cue attributes are correlated with information stored in memory as traces. The outcome of this process describes the potential activation of each trace given its association with the current cue [18, 23]. This is followed by a *conversion*-stage, in which the degree of activation determines whether the information in a trace is recollected or not [18]. This model provides a simple theoretical framework to conceptualize the influence that a situation has on the occurrence and content of episodic memories.

2.4.2. Computational Modeling of Episodic Memory Processes

Research in the domain of artificial intelligence has produced several computational models of episodic memory that implement retrieval mechanisms akin to ecphoric processing (e.g. [24–27]). A common approach is to represent both cues and traces as an array of features, and to calculate the associative strength between them using a form of distance metric. Overall, a variety of plausible models of ecphoric processing exist in the agent and cognitive modeling fields. This is of importance as it means that, when such models can be populated with actual experience-rich content from users, they can be a start to simulate their episodic memory processes. This can be combined with a data-driven approach where a model learns over time which associations are more likely to occur for a person by receiving explicit feedback from them.

2.4.3. Challenges in Ecphoric Memory Processing

Several challenges for a computational model of ecphoric processing are important to discuss here. First, in order to facilitate a useful simulation of the evocative potential of situations, it is necessary to develop a representation for them that captures their potential to act as memory cues. Developing such a representation is challenging, since it must capture attributes at different levels of abstraction, i.e. facilitate both perceptual and semantic associations.

Second, the detection of what attributes of a situation are relevant for the process of memory elicitation is a difficult and unsolved problem. The main challenge here is that a stimulus can only act as a cue in a situation if a user is actually perceiving it. So, either a system must be certain that he or she attends to it due to the context of its presentation (e.g., in the case of it taking a large amount of screen estate), or we need means to estimate the target of a user's attention (e.g. through detecting users' attentional focus via gaze-tracking, see [17])

Finally, a crucial resource required in ecphoric processing is a collection of personal information that describes those past experiences that may potentially resurface as episodic memories. In our architecture, these form the records of the Episodic Memory Store. One particular challenge here is the *comprehensiveness* required from these records: to meaningfully contribute in overcoming the RECAP problem, the they need to cover enough ground about users' lives to facilitate association with the events relevant for in a given situation. A starting point for its construction can be the substantial research on the creation of *lifelogs*. It describes the collection and organization of large quantities of data describing a person's experiences into a single comprehensive digital repository [28]. Common tasks for constructing such an archive include recording and fusing multimodal data traces into a single timeline, its automatic segmentation into a structure of distinct events [29], and its automatic semantic annotation through pattern recognition techniques [30].

While addressing policies for population and management of such an episodic store is beyond the scope of this chapter, we feel it is important to highlight the challenge of maximizing *privacy* in its construction (both of users themselves and the people they encounter in their lives). Additionally, this includes methods for providing users with control over what parts of the episodic store is available for personalization purposes. Both privacy in lifelogging [31], as well as management of long-term user models [32] are the subject of ongoing research.

2.5. COGNITIVE APPRAISAL THEORY AND THE CHALLENGE OF PREDICTING EMOTIONAL EXPERIENCE

2.5.1. PSYCHOLOGICAL BACKGROUND

A series of common evaluative judgments (e.g. novelty, goal-congruence, etc. [33]) have been identified to reliably accompany and discriminate between emotional experiences [34]. These judgments can be seen as partial mental representations of the emotional qualities of experiences [35]. The view of relating cognitive judgments of personal meaning to emotional responses is called *Cognitive Appraisal Theory (CAT)*. Its central assumption is that an organism's emotional responses express how much personal significance it assigns to the information it processes in a given situation w.r.t its utility for the fulfillment of its concerns [33]. While the descriptions of appraisal processing given in the literature often focus on evaluations of individuals' immediate surroundings, CAT argues that emotional appraisal is a fundamental mode of cognitive-affective functioning. As such, it applies to any kind of experiential content: perceived, remembered and even imagined [36]. In summary, CAT provides a general theoretical lens for understanding and describing the emergence of emotional qualities in experiences based on information processing. Because of this, we see them as a promising approach to model the relationship between episodic memories and the emotional qualities of their experience for a person.

2.5.2. COMPUTATIONAL MODELING OF EMOTIONAL APPRAISAL

Numerous computational models have drawn on appraisal theories to enable virtual agents or robots to display plausible emotional reactions to events in their artificial environments or in interactions with users (see [37] for a comprehensive overview).

In addition, artificial intelligence research has used appraisal theory to enable virtual agents to reason about the potential emotional reactions of human beings that they are interacting with (e.g. [38]). Despite the popularity of appraisal theories as inspiration for computational models of emotion elicitation, they have not seen wide usage in models of experiencing episodic memories. However, several existing computational memory models for intelligent agents include an abstract representation of their emotional state (e.g. [39]) or appraisal values [40] to describe the emotional experience that an actor has had in a previous event. This work shows that it is feasible to model the appraisal of events in a personal context. Although work on appraising the situation and memories of an actual person (rather than a virtual agent or robot) is scarce, the modeling technique can be similar.

2.5.3. CHALLENGES IN COGNITIVE APPRAISAL OF EPISODIC MEMORIES

With the exception of [26] there has been no research on computational modeling of how episodic memories are appraised upon recollection. This may be in part because there are some conceptual challenges to a straightforward application of established appraisal theories to episodic memories as stimulus events that form the target of appraisal. Especially challenging is the fact that episodic memories contain multiple aspects that can be appraised by individuals. On the one hand, there is the recollected information itself (which already has been appraised in the past during the original experience). On the other hand, there is information available that describes the current circumstances under which the event is recollected, such as its relevance for a person's current motivations. How these different sources of affective information shape the outcome of a person's emotional interpretation of a situation needs to be accounted for in a computational model of this process. For this reason it is important to investigate how common dimensions in appraisal theories can most meaningfully be applied in a computational model of episodic memories, as well as in how far such an application produces outcomes that are plausible and congruent with human experiences at the moment of recollection.

An additional challenge is the inference of a person's current goals. CAT postulates that motivations play an essential role in appraisal processing, but these constructs cannot be directly observed in individuals. As such, research contributing to their automatic inference from individuals' behavior has a tremendous potential for supporting the computational modeling of emotional appraisal processing in users. Existing technologies, such as data-driven and automatic driver intention recognition (for short term goals) [41] and explicit preference elicitation (long term values and preferences of people) [42], demonstrate that this is at least a feasible road to take. Furthermore, existing research on activity recognition [43] can be already used for coarse goal detection (going to work, going to bed, etc.). As such, there is quite some work showing that the inference of users' goals and intentions at different time scales is at least a feasible enterprise, given sufficient sensor data.

2.6. SUMMARY AND CONCLUSION

Experiences from our past are a primary influence on how we understand our environment in the present, including during interactions with multimedia applications. Ignoring

this influence results in recollection-unaware media technology that is oblivious to the RECAP problem for personalization. Ramifications become strikingly evident when looking at scenarios where the primary goal of applications hinges on their capacity to shape experiences through elicitation of episodic memories.

We have argued that providing media technologies with increased empathy for their human users requires enabling them to display awareness of when and how they dynamically experience their past in episodic memories. Our proposed architecture for an *Artificial Empathic Memory* forms a psychologically-grounded computational blueprint for providing applications with the means to do this. It comprises a series of processing components that jointly form a computational model of how externally triggered episodic remembering influences affective experience. Access to this information enables applications to adjust their behavior in meaningful ways, thereby facilitating truly personalized experiences.

Instantiating the individual components of such an AEM is a challenging task. However, it benefits from existing technological research in a variety of areas, such as the detection of attentional deployment from multimodal sensor data, computational cognitive modeling, and the development of lifelogging appliances. Given this, we feel that there is no fundamental technological hurdle for developing applications that better understand their users' subjective experiences by accounting for the role of episodic recollections in them.

REFERENCES

[1] W. Ickes, *Empathic Accuracy,* Journal of Personality **61**, 587 (1993).

[2] J. Decety and P. L. Jackson, *The functional architecture of human empathy.* Behavioral and cognitive neuroscience reviews **3**, 71 (2004).

[3] L. Stinson and W. Ickes, *Empathic Accuracy in the Interactions of Male Friends Versus Male Strangers,* Journal of Personality and Social Psychology **62**, 787 (1992).

[4] D. van Gennip, E. van den Hoven, and P. Markopoulos, *Things That Make Us Reminisce,* in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, CHI '15 (ACM, New York, NY, USA, 2015) pp. 3443–3452.

[5] M. Vannucci, C. Pelagatti, M. Hanczakowski, G. Mazzoni, and C. R. Paccani, *Why are we not flooded by involuntary autobiographical memories? Few cues are more effective than many,* Psychological Research **79**, 1077 (2014).

[6] J. Nakamura and M. Csikszentmihalyi, *The concept of flow,* in *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi* (Springer Netherlands, Dordrecht, 2014) pp. 239–263.

[7] M. Csikszentmihalyi, S. Abuhamdeh, and J. Nakamura, *Flow,* in *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi* (Springer Netherlands, Dordrecht, 2014) pp. 227–238.

[8] F. Massimini, M. Csikszentmihalyi, and M. Carli, *The monitoring of optimal experience a tool for psychiatric rehabilitation,* Journal of Nervous and Mental Disease **175**, 545 (1987).

[9] S. D'Mello, A. Graesser, and R. W. Picard, *Toward an affect-sensitive autotutor,* IEEE Intelligent Systems **22**, 53 (2007).

[10] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, *Boredom, engagement and anxiety as indicators for adaptation to difficulty in games,* in *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era - MindTrek '08*, MindTrek '08 (ACM Press, New York, New York, USA, 2008) p. 13.

[11] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, *Fusing visual and behavioral cues for modeling user experience in games,* IEEE Transactions on Cybernetics **43**, 1519 (2013).

[12] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, *Detecting engagement in hri: An exploration of social and task-based context,* in *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012* (IEEE, 2012) pp. 421–428.

[13] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang, *Predicting human interruptibility with sensors,* ACM Transactions on Computer-Human Interaction **12**, 119 (2005).

[14] L. D. Turner, S. M. Allen, and R. M. Whitaker, *Interruptibility prediction for ubiquitous systems,* in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, UbiComp '15 (ACM, New York, NY, USA, 2015) pp. 801–812.

[15] S. M. Wierda, H. van Rijn, N. A. Taatgen, and S. Martens, *Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution,* Proceedings of the National Academy of Sciences **109**, 8456 (2012).

[16] P. M. Corcoran, F. Nanu, S. Petrescu, and P. Bigioi, *Real-time eye gaze tracking for gaming design and consumer electronics systems,* IEEE Transactions on Consumer Electronics **58**, 347 (2012).

[17] C. Roda and J. Thomas, *Attention aware systems: Theories, applications, and research agenda,* Computers in Human Behavior **22**, 557 (2006).

[18] E. Tulving, *Synergistic ecphory in recall and recognition.* Canadian Journal of Psychology/Revue canadienne de psychologie **36**, 130 (1982).

[19] M. A. Conway and C. W. Pleydell-Pearce, *The construction of autobiographical memories in the self-memory system.* Psychological Review **107**, 261 (2000).

[20] S. M. Smith and E. Vela, *Environmental context-dependent memory: A review and meta-analysis,* Psychonomic Bulletin and Review **8**, 203 (2001).

[21] E. Tulving and D. M. Thomson, *Encoding specificity and retrieval processes in episodic memory,* Psychological Review **80**, 352 (1973).

[22] J. H. Mace, *Involuntary autobiographical memories are highly dependent on abstract cuing: The proustian view is incorrect,* Applied Cognitive Psychology **18**, 893 (2004).

[23] E. Tulving, M. E. L. Voi, D. A. Routh, and E. Loftus, *Ecphoric Processes in Episodic Memory [and Discussion],* Philosophical Transactions of the Royal Society B: Biological Sciences **302**, 361 (1983).

[24] A. M. Nuxoll and J. E. Laird, *Enhancing intelligent agents with episodic memory,* Cognitive Systems Research **17-18**, 34 (2012).

[25] F. Rabe and I. Wachsmuth, *An Event Metric and an Episode Metric for a Virtual Guide,* in *Proceedings of the 5th International Conference on Agents and Artificial Intelligence*, Vol. 2, edited by J. Filipe and A. Fred (SciTePress, Barcelona, Spain, 2013) pp. 543–546.

[26] P. F. Gomes, C. Martinho, and A. Paiva, *I've Been Here Before! Location and Appraisal in Memory Retrieval,* in *International Conference on Autonomous Agents and Multiagent Systems*, Aamas (2011) pp. 1039–1046.

[27] C. Brom and J. Lukavský, *Towards more human-like episodic memory for more human-like agents,* in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 5773 LNAI (2009) pp. 484–485.

[28] C. Gurrin, D. Byrne, N. O'Connor, G. Jones, and A. Smeaton, *Architecture and challenges of maintaining a large-scale, context-aware human digital memory,* in *5th International Conference on Visual Information Engineering (VIE 2008)* (2008) pp. 158–163.

[29] A. R. Doherty and A. F. Smeaton, *Automatically segmenting lifelog data into events,* in *WIAMIS 2008 - Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services* (IEEE, 2008) pp. 20–23.

[30] C. Gurrin, A. F. Smeaton, and A. R. Doherty, *LifeLogging: Personal Big Data,* Foundations and Trends® in Information Retrieval **8**, 1 (2014).

[31] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia, *Enhancing Lifelogging Privacy by Detecting Screens,* in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (ACM Press, New York, New York, USA, 2016) pp. 4309–4314.

[32] D. Barua, *A framework for user controlled remembering and forgetting in long term user models,* in *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11* (ACM Press, New York, New York, USA, 2011) p. 527.

[33] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, *Appraisal theories of emotion: State of the art and future development,* Emotion Review **5**, 119 (2013).

[34] K. R. Scherer, *The role of culture in emotion-antecedent appraisal,* Journal of Personality and Social Psychology **73**, 902 (1997).

[35] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross, *The Experience of Emotion,* Annual Review of Psychology **58**, 373 (2007).

[36] J. Broekens, D. DeGroot, and W. A. Kosters, *Formal models of appraisal: Theory, specification, and computational model,* Cognitive Systems Research **9**, 173 (2008).

[37] S. Marsella, J. Gratch, and P. Petta, *Computational models of emotion,* in *A Blueprint for Affective Computing - A sourcebook and manual*, edited by E. Scherer, K., Bänziger, T., Roesch (Oxford University Press, 2010) pp. 21–41.

[38] H. Boukricha, I. Wachsmuth, M. N. Carminati, and P. Knoeferle, *A computational model of empathy: Empirical evaluation,* in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013* (IEEE, 2013) pp. 1–6.

[39] C. Brom, K. Pešková, and J. Lukavský, *What Does Your Actor Remember? Towards Characters with a Full Episodic Memory,* in *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*, edited by M. Cavazza and S. Donikian (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007) pp. 89–101.

[40] J. Dias, W. C. Ho, T. Vogt, N. Beeckman, A. Paiva, and E. André, *I Know What I Did Last Summer: Autobiographic Memory in Synthetic Characters,* in *Affective Computing and Intelligent Interaction*, edited by A. C. R. Paiva, R. Prada, and R. W. Picard (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007) pp. 606–617.

[41] A. Doshi and M. M. Trivedi, *Tactical driver behavior prediction and intent inference: A review,* in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* (IEEE, 2011) pp. 1892–1897.

[42] L. Chen and P. Pu, *EPFL Technical Report*, Tech. Rep. (Ecole Politechnique Federale de Lausanne (EPFL), IC/2004/67, 2004).

[43] L. Bao and S. S. Intille, *Activity Recognition from User-Annotated Acceleration Data,* in *Pervasive Computing: Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004. Proceedings*, edited by A. Ferscha and F. Mattern (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004) pp. 1–17.

3

Collecting the Mementos Dataset: A Multimodal Corpus of Affect and Memory Processing in Response to Video Stimuli

This chapter is based on: Dudzik, B., Hung, H., Neerincx, M. A., & Broekens, J. (2021). *Collecting Mementos: A Multimodal Dataset for Context-Sensitive Modeling of Affect and Memory Processing in Responses to Videos.* IEEE Transactions on Affective Computing, 3045(c), 1–1. https://doi.org/10.1109/TAFFC.2021.3089584

ABSTRACT

In this chapter 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* 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.

3.1. 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, 4].

While existing research has primarily focused on predicting the immediate emotional impact of video viewing on individuals [5], efforts have also touched on the ebb and flow of viewers' attention while doing so [6], or the ability of content to be remembered [7]. 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 [8].

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 [9–11]. 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.
- **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.

3.2. 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 [12], and the ability of such recollections to have a substantial emotional impact [13]. Moreover, findings suggest that the emotional impact of media stimuli on individuals matches their feelings towards the connected memories [14]. 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., [15]). 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 [6]. 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. [16] 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.

3.2.1. 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 [17]. Such differences in context are known to strongly affect both emotional experience and expression in general [18], particularly in response to media content [8]. Similar findings exist for the influence of context on the elicitation of personal memories [19]. 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 [8]. 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., [20–22], 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* [23]), as well as in databases for affect modeling in computer science (e.g. *DEAP* [24], and *AMIGOS* [25]). 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 [8]). 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

Table 3.1: Comparison of Databases for Video-Induced Affect

Database		Data Collection		Videos Stimuli		Sensor Data		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 40 participants in a Lab

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.

3.2.2. 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 [32]. 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 [33]). Similarly, people may respond differently to video content, depending on the cultural values of the country that they are nationals of [34]. Moreover, personality traits provide broad insights into individual differences between people and explain variation in affective responses to videos [34]. In contrast to emotions, moods are enduring, low-intensity affective states that are typically not directed towards a specific event or stimulus [35]. Nevertheless, they can exercise a broad influence on individuals' experience and behavior in a given situation, including affective responses to videos [8].

3.2.3. 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 [36]. 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 [37], and 2. the accurate representation of psychological and social constructs [36]. 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. In particular, we measure individuals' affective responses in terms of the widely used *Pleasure-Arousal-Dominance (PAD)* framework [38], using the AffectButton, a well-validated measurement instrument [39]. 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 [40]).

3.2.4. RELATED WORK

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 3.1 for an overview). For this purpose, we differentiate between corpora that are either 1. *Stimulus-Centric (SC)*, or 2. *Viewer-Centric (VC)*, depending primary 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 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 [5]. Additionally, affect is often labeled at the video level, e.g., through aggregating ratings for the same stimulus. Principal examples include *LIRIS-ACCEDE* [20], or *VideoEmotion* [30]. In contrast, corpora focusing on viewers rely on a comparatively small set of videos for emotion induction to capture self-reports and multimodal measures from a larger pool of individuals. They are primarily used for work on *Multimodal Affect Detection* [41], i.e., analyzing behavioral, physiological, and sometimes contextual data to predict the emotional response of individuals. Relevant examples include *DEAP*[24], *DECAF*[28], *MAHNOB-HCI*[29], *AMIGOS* [25], *ASCERTAIN* [26], and *SEWA* [31]. Noteworthy is also *CP-QAE-I* [27], which does not contain behavioral measures, but provides rich context about individual viewers' background. Importantly, SC 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, it does not offer physiological measures for analysis, unlike them. 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 media stimuli.

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. [42] and the dataset used for the Memorability-task at MediaEval. [43]. 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 [6]. 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.

3.3. 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.

3.3.1. 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* [44]. The idea behind this alignment is to maximize the capacity of our stimuli to trigger personal memories in viewers.

3.3.2. MEASURES AND MATERIALS

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* [24]. 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. [13]). Secondly, 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 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 3.2* for a description of the selected video stimuli, including the title, release year, and genre.

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)* [45], 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

Table 3.2: 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

the *AffectButton* [39] 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 [39] 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.

- **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"*.

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.

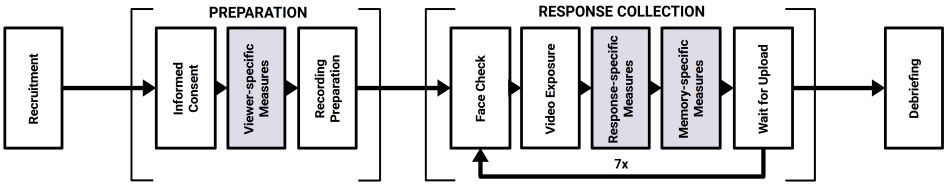


Figure 3.1: Protocol for data collection in our online study from participating crowd-workers. Purple fields refer to stages at which we collected the respective Self-report Measures listed in *Section 3.2*.

3

ONLINE APPLICATION FOR DATA COLLECTION

For collecting data from participants we developed a specific online application based around the JavaScript-framework *jsPsych*[1] [46], 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.

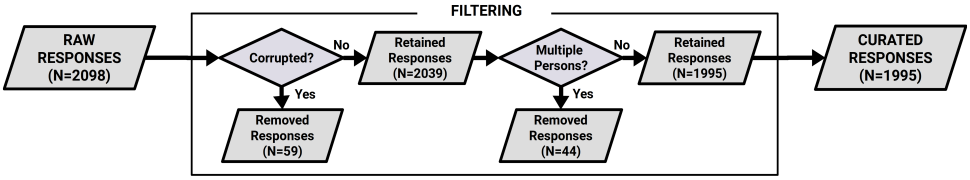


Figure 3.2: Filtering of Invalid Responses from the Dataset.

3.3.3. PROTOCOL

In the following, we describe the different phases of the protocol of our online study for data collection. See *Figure 3.1* for a graphical overview.

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

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.

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.

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.

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.

3.3.4. ETHICS STATEMENT

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

3.4. 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]

3.4.1. CURATION

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 [47], we represent intensity as the magnitude of ratings in terms of PAD-scores, using the following formula:

$$I = \frac{1}{\sqrt{3}} \sqrt{\left(p^2 + ((a+1)/2)^2 + d^2\right)} \tag{3.1}$$

  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,

[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.

motivated by the layout of the AffectButton instrument, which maps maximum negative arousal to neutral face representations in the centre of the widget [39].

- **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 [48]. 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* [49]. 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 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.

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 *Figure 3.2*.

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] [50] 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.

3.4.2. 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 3.3* for summary

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

statistics).

Table 3.3: Statistics of Self-Report Data for Responses to Videos (Curated Dataset)

Variable	Measure	M (SD)	Min/Max
Personality	**Hon/Hum.**	2.67 (0.75)	0.5/4.0
$N = 297^{+}$	**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	**Pleasure**	0.42 (0.4)	-0.75/1.0
$N = 297^{+}$	**Arousal**	-0.12 (0.77)	-1.0/1.0
	Dominance	0.38 (0.47)	-1.0/1.0
	Intensity		
Demographics	**Age**	33.06 (6.01)	25.0/46.0
$N = 297^{+}$		$Unique$	Top ($Freq$)
	Nationality	3	USA (240)
	Gender	2	male (159)
Ind. Emotion	**Pleasure**	0.2 (0.53)	-1.0/1.0
$N = 1995^{*}$	**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	**Prev. Expo.**	0.26 (0.16)	0.2/1.0
$N = 1995^{*}$			
Mem. Content	**Descr. (WC)**	22.61 (13.38)	2/89
$N = 989^{\dagger}$	**Descr. (FRES)**	78.0 (16.77)	-8.73/119.19
		$Unique$	Top ($Freq$)
	Age in Mem.	5	11-20y (437)
Mem. Affect	**Pleasure**	0.33 (0.53)	-1.0/1.0
$N = 989^{\dagger}$	**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.	**Vividn.**	0.64 (0.27)	0.01/1.0
$N = 989^{\dagger}$	**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.

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 [39] for a discussion).

RESPONSE-SPECIFIC MEASURES

Induced Emotion: Similarly to the mood scores of participants, their emotional responses to the videos tend to be 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 ($M(SD) = .26(0.16)$).

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% of all recollections ($N = 52$) involved more than 2 of them.

Memory Content: In *Figure 3.4* 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: $M(SD) = 22.83(13.55)$),

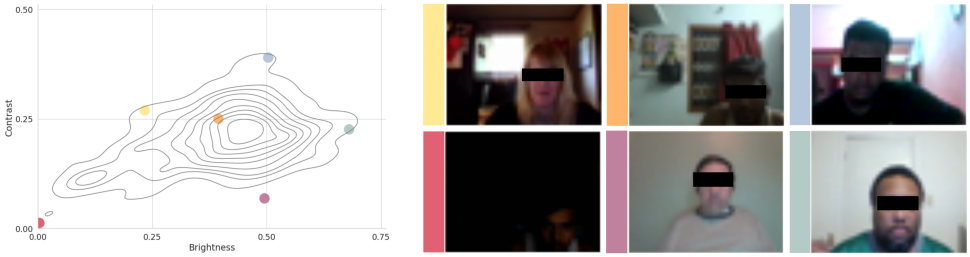


Figure 3.3: Contour plot of the average brightness and contrast of the frames in recordings collected from participants. Images on the right are example frames taken from the recordings at the marked locations (down-sampled and masked to preserve participants privacy).

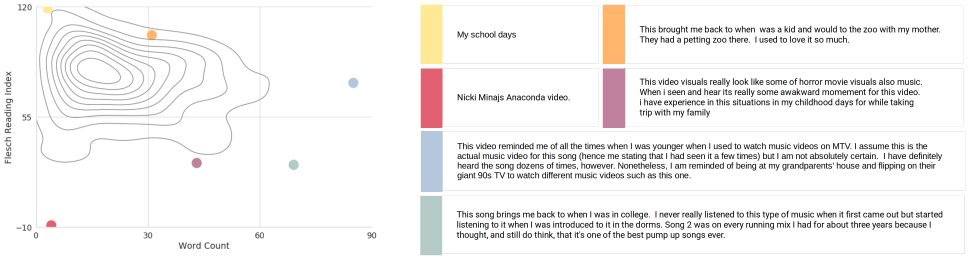


Figure 3.4: Contour plot of the Word Count and Flesh Reading-Ease Score (FRES) of the memory descriptions collected from participants. Text fragments on the right are examples from the marked locations.

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 ($M(SD) = 0.55(0.32)$.

3.4.3. STATISTICS RECORDED BEHAVIOR

Here we provide a brief overview and discussion of the behavioral recordings captured from participants (see *Table 3.4* for summary statistics).

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

Table 3.4: 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;

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 *Figure 3.3*, for a visual impression of this diversity).

FACIAL EXPRESSIONS

OpenFace detected faces successfully as present in most of the frames (99% 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).

3.5. ANALYSIS OF VALIDITY

3.5.1. 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 *Figure 3.5*). 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

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

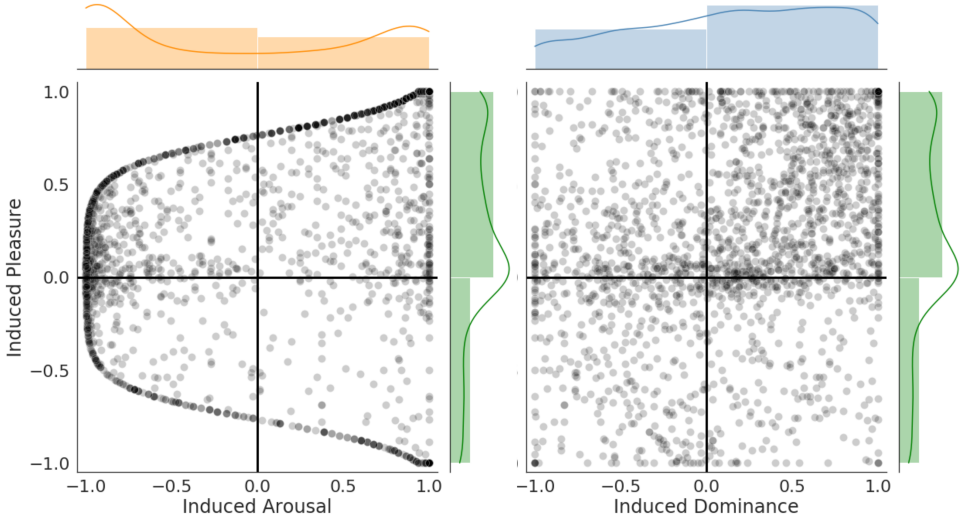


Figure 3.5: Distribution of **Induced Emotion for Individual Responses** in the Pleasure-Arousal and Pleasure-Dominance Planes ($N = 1995$).

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 *Figure 3.6*). 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% ($N = 606$) 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 [51]. 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 use the memories contained in Mementos to study primarily or model differences between neutral and positive associations of pleasure and dominance.

3.5.2. 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

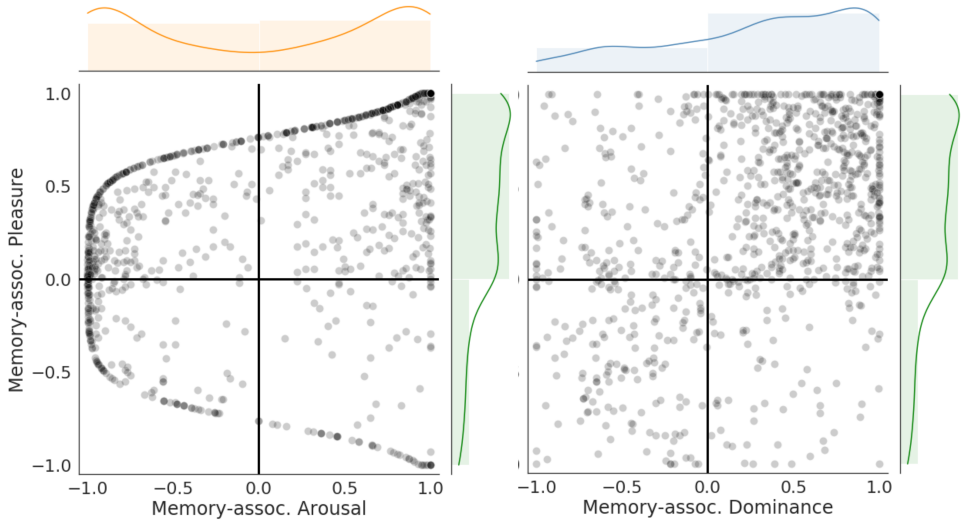


Figure 3.6: Distribution of **Associated-Affect for Individual Memories** in the Pleasure-Arousal and Pleasure-Dominance Planes ($N = 989$).

are presented in greater detail in other publications using the Mementos dataset [9–11][6].

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% of the total variation in responses, leaving the remaining 86% 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 *Figure 3.7*. 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

[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.

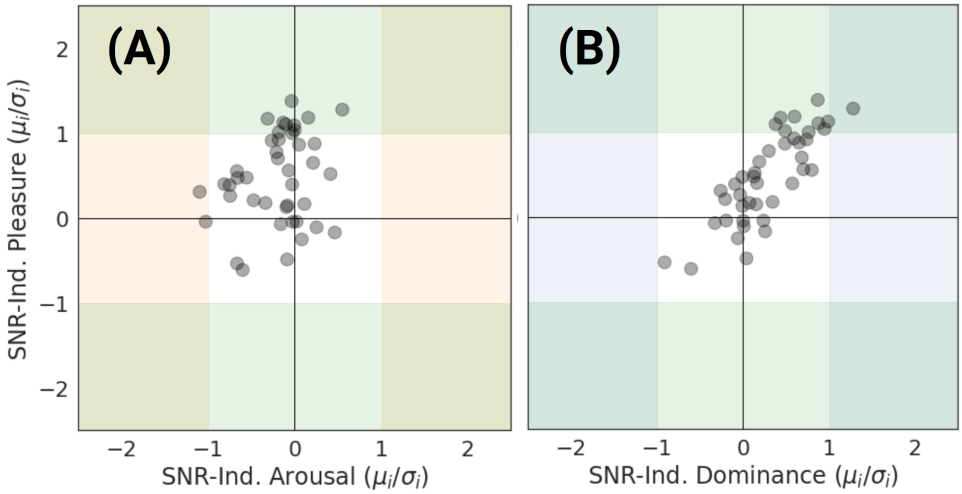


Figure 3.7: Plots of the video-wise Signal-to-Noise Ratio (*SNR*) Score for ratings of **Induced Emotions** on *(A)* the pleasure-arousal plane and *(B)* pleasure-dominance plane. Colors demarcate regions where the mean ratings for a video on the respective axis exceed its standard deviation.

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. [9]. 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 [14, 52]. Overall, they point to the validity of Mementos as a corpus capturing interactions between personal memories and affective processing.

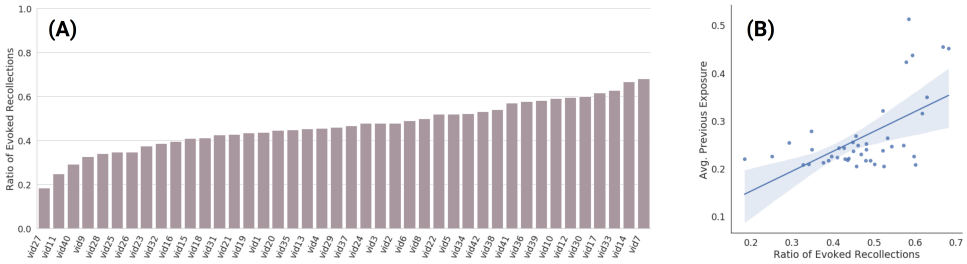


Figure 3.8: **(A)** Video-wise rates at which exposure evoked a recollections (i.e. at least one personal memory is triggered). **(B)** Scatter plot with linear relationship between the video-wise Rates of Evoked Recollections and Viewers' Average Degree of Previous Exposure to them. Shaded area denotes the 95% confidence interval.

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. [9] for the detailed analysis). In particular, we find that viewers' personality does not have a significant effect on their induced emotions under these circumstances, and differences in demographics and mood only have a small impact (Demographics: $Avg\Delta R^2_m = .013$, Mood: $Avg\Delta R^2_m = .014$). 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% of the variance across induced pleasure ($F(13,223.33) = 2.14$, $p < .05$, $R^2_m = 0.032$), arousal ($F(13,232.87) = 3.12$, $p < .001$, $R^2_m = 0.06$) and dominance ($F(13,232.48) = 2.94$, $p < .001$, $R^2_m = 0.057$). Together, this shows that Mementos captures individual differences and mood effects, mirroring findings in other research on responses to video content (e.g., [8, 34]), thereby adding to the validity of the corpus.

MEMORY EVOCATION

Effects of Video Stimuli and Familiarity: Previous findings indicate that video stimuli can substantially differ in their capacity to trigger personal memories [12]. In particular, for musical material, one variable associated with its evocative potential is familiarity with it [13]. 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 *Figure 3.8-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 *Figure 3.8-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.

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 [53]. 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* [54]. 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% 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.

3.5.3. ANALYSIS OF MULTIMODAL DATA

WEBCAM RECORDINGS

Impact of Lighting Conditions on Facial Analysis: Lighting conditions can pose a challenge for vision-based face analysis [55]. 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 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.

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 detail 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 [10]. 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 [10, 11] (see also *Section 3.6* 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. Especially, 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.

3.6. 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.

3.6.1. CONTEXT-SENSITIVE VIDEO AFFECTIVE CONTENT ANALYSIS [8]

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 [10]. 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 3.5.3*). 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,

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

[8] Note that the work summarized here is described in detail as part of this dissertation in *Chapter 5*

despite the increased complexity of implementation.

3.6.2. USE FOR AFFECTIVE BEHAVIOR ANALYSIS [9]

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 [56] and limit performance in real-world scenarios. For this reason, we explore the potential of automatically analyzing video and memory alongside facial behavior to support affect detection in another article, using the Mementos dataset [11]. Besides extracting distinct feature sets for representing video content and descriptions of viewers' personal memories, we use 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 show that adding context provides overall performance improvements, they also shed insights into the complementary nature of affective information sources. Notably, facial expressions offer unique benefits for predicting arousal, while video and memory content explain 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 conditions and potentially sparse facial expressiveness from participants.

3.7. POTENTIAL USES IN FUTURE RESEARCH

3.7.1. 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 3.6*: 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 3.2.4*), 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 [11], or they may focus more extensively on the other contextual factors it

[9] Note that the work summarized here is described in detail as part of this dissertation in *Chapter 6*

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 [5], as well as Soleymani et al. [8] 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).

3.7.2. 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 [57]. In particular, information about the affect associated with past experiences involving products and services might be important for designing and personalizing these [58]. 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 [59]. 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.

3.7.3. 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 [60]) or even disastrous consequences (e.g., a distracted driver on the road [61]). The recollection of personal memories can require significant internal attentional engagement [62]. As such, these may result in behavioral responses similar to other types of internal cognitive processing, which can be detected through gaze behavior [63]. 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 [9]. 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.

3.8. 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 [64]). 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. [65]

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 [31]) 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 [66], 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. [12]). 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. [67]).

3.9. 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.

REFERENCES

[1] A. Bartsch, *Emotional Gratification in Entertainment Experience. Why Viewers of Movies and Television Series Find it Rewarding to Experience Emotions,* Media Psychology **15**, 267 (2012).

[2] S. Wang and Q. Ji, *Video affective content analysis: A survey of state-of-the-art methods,* IEEE Transactions on Affective Computing **6**, 410 (2015).

[3] A. Hanjalic, *Extracting moods from pictures and sounds: towards truly personalized TV,* IEEE Signal Processing Magazine **23**, 90 (2006).

[4] M. Soleymani and M. Pantic, *Emotionally Aware TV,* Proceedings of TVUX-2013: Workshop on Exploring and Enhancing the User Experience for TV at ACM CHI 2013 (2013).

[5] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, *Affective video content analysis: A multidisciplinary insight,* IEEE Transactions on Affective Computing **9**, 396 (2018).

[6] A. Stewart, N. Bosch, H. Chen, P. Donnelly, and S. D'Mello, *Face Forward: Detecting Mind Wandering from Video During Narrative Film Comprehension,* in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10331 LNAI (Springer, Cham, 2017) pp. 359–370.

[7] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, and A. Oliva, *Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability,* (2020), arXiv:2009.02568 .

[8] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, *Corpus development for affective video indexing,* IEEE Transactions on Multimedia **16**, 1075 (2014), arXiv:1211.5492 .

[9] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, *Investigating the Influence of Personal Memories on Video-Induced Emotions,* in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (ACM, New York, NY, USA, 2020) pp. 53–61.

[10] 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 preprint, arXiv:2008.12096, arXiv:2008.12096 [cs.HC] .

[11] 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 *Proceedings of the 2020 International Conference on Multimodal Interaction*, Vol. 10 (ACM, New York, NY, USA, 2020) pp. 153–162.

[12] D. G. McDonald, M. A. Sarge, S.-F. Lin, J. G. Collier, and B. Potocki, *A Role for the Self,* Communication Research **42**, 3 (2015).

[13] P. Janata, S. T. Tomic, and S. K. Rakowski, *Characterisation of music-evoked autobiographical memories,* Memory **15**, 845 (2007).

[14] H. Baumgartner, M. Sujan, and J. R. Bettman, *Autobiographical Memories, Affect, and Consumer Information Processing,* Journal of Consumer Psychology **1**, 53 (1992).

[15] E. van den Hoven, *A future-proof past: Designing for remembering experiences,* Memory Studies **7**, 370 (2014).

[16] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, *Artificial Empathic Memory,* in *Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions - EE-USAD'18* (ACM Press, New York, New York, USA, 2018) pp. 1–8.

[17] C. Lagger, M. Lux, and O. Marques, *What Makes People Watch Online Videos,* Computers in Entertainment **15**, 1 (2017).

[18] K. H. Greenaway, E. K. Kalokerinos, and L. A. Williams, *Context is Everything (in Emotion Research),* Social and Personality Psychology Compass **12**, e12393 (2018).

[19] S. M. Smith and E. Vela, *Environmental context-dependent memory: A review and meta-analysis,* Psychonomic Bulletin & Review **8**, 203 (2001).

[20] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, *LIRIS-ACCEDE: A video database for affective content analysis,* IEEE Transactions on Affective Computing **6**, 43 (2015).

[21] D. McDuff and M. Soleymani, *Large-scale Affective Content Analysis: Combining Media Content Features and Facial Reactions,* in *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge* (IEEE, 2017) pp. 339–345.

[22] J. J. Sun, T. Liu, A. S. Cowen, F. Schroff, H. Adam, and G. Prasad, *EEV Dataset: Predicting Expressions Evoked by Diverse Videos,* (2020), arXiv:2001.05488 .

[23] S. Carvalho, J. Leite, S. Galdo-Álvarez, and Ó. F. Gonçalves, *The emotional movie database (EMDB): A self-report and psychophysiological study,* Applied Psychophysiology Biofeedback **37**, 279 (2012).

[24] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, C. Mühl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *DEAP: A Database for Emotion Analysis Using Physiological Signals,* IEEE Transactions on Affective Computing **3**, 18 (2012).

[25] J. A. Miranda Correa, M. K. Abadi, N. Sebe, and I. Patras, *AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups,* IEEE Transactions on Affective Computing **39**, 1 (2018).

[26] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, *ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors,* IEEE Transactions on Affective Computing **9**, 147 (2018).

[27] S. C. Guntuku, M. J. Scott, Huan Yang, G. Ghinea, and Weisi Lin, *The CP-QAE-I: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia,* in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)* (IEEE, 2015) pp. 1–7.

[28] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, *DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses,* IEEE Transactions on Affective Computing **6**, 209 (2015).

[29] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *A multimodal database for affect recognition and implicit tagging,* IEEE Transactions on Affective Computing **3**, 42 (2012).

[30] Y.-G. Jiang, B. Xu, and X. Xue, *Predicting Emotions in User-Generated Videos,* AAAI , 73 (2014).

[31] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, *SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild,* IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 1022 (2021), arXiv:1901.02839 .

[32] B. Dudzik, M.-P. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong, *Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases,* in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 206–212.

[33] 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 **13**, 212 (2012).

[34] 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 *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015* (IEEE, 2015) pp. 236–242.

[35] K. R. Scherer, *What are emotions? And how can they be measured?* Social Science Information **44**, 695 (2005).

[36] J. Broekens, T. Bosse, and S. C. Marsella, *Challenges in Computational Modeling of Affective Processes,* IEEE Transactions on Affective Computing **4**, 242 (2013).

[37] J. A. Allen, C. Fisher, M. Chetouani, M. M. Chiu, H. Gunes, M. Mehu, and H. Hung, *Comparing Social Science and Computer Science Workflow Processes for Studying Group Interactions,* Small Group Research **48**, 568 (2017).

[38] A. Mehrabian, *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament,* Current Psychology **14**, 261 (1996).

[39] J. Broekens and W.-P. Brinkman, *AffectButton: A method for reliable and valid affective self-report,* International Journal of Human-Computer Studies **71**, 641 (2013).

[40] P. J. Lang, M. M. Bradley, B. N. Cuthbert, and Others, *International affective picture system (IAPS): Technical manual and affective ratings,* NIMH Center for the Study of Emotion and Attention **1**, 39 (1997).

[41] S. K. D'mello and J. Kory, *A Review and Meta-Analysis of Multimodal Affect Detection Systems,* ACM Computing Surveys **47**, 1 (2015).

[42] R. Samide, R. A. Cooper, and M. Ritchey, *A database of news videos for investigating the dynamics of emotion and memory,* Behavior Research Methods **52**, 1469 (2020).

[43] R. Cohendet, C.-H. Demarty, N. Q. D. K., M. Sjöberg, B. Ionescu, and T.-T. Do., *Mediaeval 2018: Predicting media memorability task.* in *Proceedings of the MediaEval 2017 Workshop* (2018).

[44] M. A. Conway and S. Haque, *Overshadowing the Reminiscence Bump: Memories of a Struggle for Independence,* Journal of Adult Development **6**, 35 (1999).

[45] R. E. De Vries, *The 24-item Brief HEXACO Inventory (BHI),* Journal of Research in Personality **47**, 871 (2013).

[46] J. R. de Leeuw, *jsPsych: A JavaScript library for creating behavioral experiments in a Web browser,* Behavior Research Methods **47**, 1 (2015).

[47] R. Reisenzein, *Pleasure-arousal theory and the intensity of emotions.* Journal of Personality and Social Psychology **67**, 525 (1994).

[48] R. Flesch, *A new readability yardstick,* Journal of Applied Psychology **32**, 221 (1948).

[49] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, *OpenFace 2.0: Facial Behavior Analysis Toolkit,* in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (IEEE, 2018) pp. 59–66.

[50] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,* IEEE Transactions on Pattern Analysis and Machine Intelligence , 1 (2019).

[51] W. R. Walker, R. J. Vogl, and C. P. Thompson, *Autobiographical memory: unpleasantness fades faster than pleasantness over time,* Applied Cognitive Psychology **11**, 399 (1997).

[52] H. Baumgartner, *Remembrance of things past: Music, autobiographical memory, and emotion,* Advances in Consumer Research **19**, 613 (1992), arXiv:9301106587 .

[53] J. M. Talarico, K. S. LaBar, and D. C. Rubin, *Emotional intensity predicts autobiographical memory experience,* Memory & Cognition **32**, 1118 (2004).

[54] G. E. Matt, C. Vázquez, and W. K. Campbell, *Mood-congruent recall of affectively toned stimuli: A meta-analytic review,* Clinical Psychology Review **12**, 227 (1992).

[55] R. Gopalan and D. Jacobs, *Comparing and combining lighting insensitive approaches for face recognition,* Computer Vision and Image Understanding **114**, 135 (2010).

[56] 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,* Psychological Science in the Public Interest **20**, 1 (2019).

[57] D. Wirtz, J. Kruger, C. N. Scollon, and E. Diener, *What to Do on Spring Break?* Psychological Science **14**, 520 (2003).

[58] D. A. Norman, *THE WAY I SEE ITMemory is more important than actuality,* Interactions **16**, 24 (2009).

[59] S. Mihailova and L. Jobson, *Association between intrusive negative autobiographical memories and depression: A meta-analytic investigation,* Clinical Psychology & Psychotherapy **25**, 509 (2018).

[60] F. Putze, D. Küster, S. Annerer-Walcher, and M. Benedek, *Dozing Off or Thinking Hard?* in *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18* (ACM Press, New York, New York, USA, 2018) pp. 258–262.

[61] 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,* Frontiers in Human Neuroscience **11**, 406 (2017).

[62] M. M. Chun, J. D. Golomb, and N. B. Turk-Browne, *A Taxonomy of external and internal attention,* Annual Review of Psychology **62**, 73 (2011).

[63] M. Benedek, R. Stoiser, S. Walcher, and C. Körner, *Eye Behavior Associated with Internally versus Externally Directed Cognition,* Frontiers in Psychology **8**, 1092 (2017).

[64] J. J. Igartua, *Identification with characters and narrative persuasion through fictional feature films,* Communications **35**, 347 (2010).

[65] S. T. Peesapati, V. Schwanda, J. Schultz, M. Lepage, S.-y. Jeong, and D. Cosley, *Pensieve,* in *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (ACM Press, New York, New York, USA, 2010) p. 2027.

[66] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: Synthetic Minority Over-sampling Technique,* Journal of Artificial Intelligence Research **16**, 321 (2002), arXiv:1106.1813 .

[67] 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 *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 697–703.

4

The Influence of Personal Memories on Video-Induced Emotions

This chapter is based on: Dudzik, B., Hung, H., Neerincx, M., & Broekens, J. (2020). Investigating the Influence of Personal Memories on Video-Induced Emotions. Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 53–61.

ABSTRACT

Making accurate predictions of he subjective emotional experience that audio-visual media content induces in individual viewers is a challenging task because of their highly person-dependent and situation-specific nature. Findings from psychology indicate that an important driver for the emotional impact of media is the triggering of personal memories in observers. However, existing research on automated predictions focuses on the isolated analysis of audiovisual content, ignoring such contextual influences. In a series of empirical investigations, we (1) quantify the impact of associated personal memories on viewers' emotional responses to music videos in-the-wild and (2) assess the potential value of information about triggered memories for personalizing automatic predictions in this setting. Our findings indicate that the occurrence of memories intensifies emotional responses to videos. Moreover, information about viewers' memory response explains more variation in video-induced emotions than either the identity of videos or relevant viewer-characteristics (e.g. personality or mood). We discuss the implications of these results for existing approaches to automated predictions and describe ways for progress towards developing memory-sensitive alternatives.

4.1. INTRODUCTION

Research on *Video Affective Content Analysis (VACA)* strives to enable technologies to automatically estimate the emotional responses videos induce in their viewers [1], e.g. to support affect-based recommendations [2, 3]. A fundamental challenge for this undertaking is that emotional experiences are highly subjective, expressing a dynamic relationship between individuals' ongoing personal needs and the perceived ability of their current situation to meet them [4]. Therefore, how people experience media stimuli depends on who they are and under what circumstances they encounter the stimuli (see, e.g. the findings of age-related differences in [5] or the existence of mood-effects [6]).

Throughout the existence of VACA research, the majority of technological investigations have avoided dealing with the issue of subjectivity. Instead, it has focused on the de-contextualized analysis of videos' audiovisual content to estimate emotional responses elicited in a majority of viewers [1]. As such, the emotional impact estimated by these technologies is not the subjective experience of a particular person viewing a video, but rather the expected response across a population of viewers, and independent of their viewing situation. However, without the capacity to reflect variations within and across individuals' impressions of the same video, the practical value of this approach for applications seems limited.

Only recently, research has started to openly address the issue of modelling individual affective experiences by exploring the usefulness of viewer- and situation-specific information in predictions, e.g. personality traits and cultural background [7], or the social setting in which viewing takes place [8]. Nevertheless, systematic research into such context-sensitive predictions remains scarce [1], and many essential drivers for human affective experiences remain unexplored and unaccounted for in computational models of video-induced emotions.

One significant influence that research has not yet touched upon is the recollection of personal memories [9]. Not only can memories about one's past have a significant emo-

tional impact in their own right, but auditory and visual material can readily trigger them in an audience [10]. Moreover, many patterns of media creation and consumption revolve specifically around this ability to serve as cues for emotionally significant memories, e.g. the taking of photos as mementoes for the future, or the listening to music from a specific period in ones' past for the sake of reminiscence and nostalgia.

In this article, we present several empirical investigations to quantify the importance of accounting for viewers' memories when predicting video-induced emotions. In particular, we present the following contributions:

(1) the collection of a dataset about viewers' recollection processes in emotional responses to music videos,

(2) a series of analyses to quantify the influence of personal memories on video-induced emotions,

(3) and a comparison of the impact on affective response prediction accuracy between information about *a)* the memories triggered in viewers, *b)* the eliciting video, and *c)* relevant viewer-characteristics (demographics, personality, and mood).

While the analyses that we present here focus on self-reported data, we discuss the relevance of our findings for existing computational approaches to predict video-induced emotions. In particular, we outline opportunities that future technological research could explore to account for the influence of personal memories in predictions made by automated systems.

4.2. BACKGROUND AND RELATED WORK

4.2.1. AFFECTIVE MEMORY PROCESSES AND MEDIA-INDUCED EMOTIONS

While individuals can intentionally recall moments from their personal history, stimuli in the external environment can also remind them of these moments involuntarily [11]. Audiovisual media appear to be potent triggers for personal memories in observers [12, 13], and these can evoke strong emotions upon recollection [10]. The mechanisms that determine when external stimuli evoke memories from the past in a person are not entirely understood. Two discovered conditions are (1) the existence of some form of semantic of perceptual association between the triggering stimulus and the recollected content [14], and (2) a state of low attentional engagement (see, e.g. [15]). Once recollection takes place, memories can elicit strong emotional responses, as is demonstrated by their frequent use in emotion induction procedures [16]. Moreover, empirical evidence points to a direct connection between the affect associated with the memories triggered by a media stimulus – i.e. how one feels about the memory – to that stimulus' emotional impact [17]. Together these findings indicate that recollection of personal memories can drive viewers' experiences.

4.2.2. VIDEO-AFFECTIVE CONTENT ANALYSIS

Video-Affective Content Analysis is an approach to predicting the emotional response of viewers to audiovisual media content. Research on this topic comprises roughly three components: the type of emotional response that forms the target of the prediction,

the granularity at which responses' to video are to be predicted, and the sources of information that form the basis for predictions.

The types of emotional responses of primary concern in VACA research take two different forms [1]. The first is the expected emotion for a video (What feelings does it evoke in the majority of viewers?). The second, is the induced emotion (What feelings does a video evoke in a particular viewer?). Regarding the granularity of predictions, research efforts comprise of those undertaking a global analysis – which attempts to predict the emotional response for an entire video –, or those conducting a continuous analysis – which consists of predictions for smaller windows (potentially down to the frame-level) [1]. Finally, existing VACA approaches are distinguishable by the information that they use to make their predictions [18]. *Direct VACA* focuses on the analysis of the audiovisual signals comprising the content of a video. In contrast, *Implicit VACA* relies on the automatic analysis of viewers' behaviour during exposure to a video clip, e.g. facial expressions (e.g. [19]) or physiological response (e.g. [20].

Overall, the majority of existing research on predicting video-induced emotions focuses on expected emotions at a global level, using *Direct VACA* [1, 18]. Because this task focuses on assigning a single expected emotion to a video, most of the existing VACA research (either implicitly or explicitly) assumes homogeneous emotional responses across viewers. Video corpora used in VACA research are specially designed to adhere to this low-variation assumption by filtering out videos that elicit diverse responses (e.g. [19, 20]). Researchers in the VACA community are aware that technologies built on this assumption cannot reflect the natural variation in viewers' responses, and that doing so requires increasingly context-sensitive approaches [1, 6, 18]. However, explorations in this regard are only recently gaining traction. Initial efforts have touched on the value of information about personality and cultural background for personalizing predictions [7, 21], or the role of viewers' overall affective mood [6]. In principle, the combination of behavioural and physiological signals of viewers in combination with features describing the video content (e.g. [2]) can also be considered as a form of contextualized prediction. However, this response information is only available when viewers' are already exposed to a video, potentially ruling out some primary use-cases for undertaking VACA in the first place, e.g. intelligent recommendation.

In summary, few prior works investigate contextual features for predictions of individual viewers' emotional responses. Existing explorations revolve primarily around addressing viewer-specific differences, with some consensus among scholars for the importance of basic demographic factors, personality traits, and mood as features. Incorporating information about viewers' cognitive processes in general, and the effects of recollecting associated personal memories, in particular, has not been explored so far.

4.2.3. REPRESENTATIONS OF EMOTIONS

There exist a wide variety of schemes for describing and classifying emotional experiences in psychology, many of which have been used in modelling video-induced emotions [1]. Broadly, these consist of two distinct groups: categorical and dimensional schemes. Categorical schemes typically build on psychological theories that assume a set of discrete emotional states, often with unique physiological response components attached to them (e.g. facial expressions, or the patterns of activity in the autonomous nervous system [22]).

In contrast, dimensional schemes describe affect as points in continuous space with a set of orthogonal dimensions (e.g. [23]). Each dimension is supposed to capture an essential quality for discriminating between affective states.

Overall, there is no consensus about the merit of any particular type of scheme among researchers engaging in modelling video-induced emotions. However, categorical approaches may lack the necessary nuance for describing affective media experiences [1]. For this reason, we have used the dimensional PAD-framework [24] to measure video-induced emotions in the dataset collected for this research (see Section 3 below). It is prominent in psychological studies and also extensively used in VACA research (e.g. [20]). PAD characterizes affective experiences along three orthogonal axes, each with a positive and negative polarity: *pleasure (P)* (valence, is an experience positive or negative, enjoyable or unpleasant?), *arousal (A)* (Does it involve a high or low degree of bodily activation or alertness?), and *dominance (D)* (To what degree am I in control of the experienced situation?).

4.3. DATA COLLECTION

In this section, we provide a detailed description of the relevant elements of a dataset that we have collected to investigate the influence of recollection processes in emotional responses to videos.

4.3.1. SELECTED VIDEO STIMULI

We used music video segments that have previously been evaluated for their affective impact in the *DEAP* dataset [20]. We opted for this choice, because (1) prior research demonstrated the potency of musical material for triggering personal recollections [10], and (2) each stimulus in this list was rated by multiple viewers using the PAD-framework. This second property provided us with information about the expected distribution of emotional responses from viewers, which we used for balancing the distribution of selected videos in the design of our study.

From the total of 150 video segments for which the validation study for the DEAP corpus collected ratings, we selected a subset of 42 videos based on their variation. In particular, we selected an equal amount of stimuli per affective dimension that possess either a high- or a low- degree of variation in their emotional responses. For example, we balanced videos where different viewers show very similar pleasure-responses with other videos where this is not the case. We opted for this scheme because we hypothesized that situation- and viewer-specific influences are more likely present in high variation stimuli. Consequently, these differences in variation might also reflect variation in the occurrence and influence of personal memories. It is important to note that for the DEAP corpus, only videos with a narrow distribution and pronounced average emotional responses were kept in their experiments [20].

4.3.2. PARTICIPANTS

We recruited 300 individuals via the crowd-sourcing platform *Amazon Mechanical Turk* and compensated each for their participation with 6 *USD*. We did not constrain our recruitment efforts to a geographic region or particular nationalities. However, we requested that individuals have command of the English language and that they participate

in an environment that allows them to pay attention for the entire duration of the task. Additionally, we enforced restrictions on the age of participants in our online survey so that they were between 25 and 46 years of age. We implemented this to ensure that the release dates of music videos used in our study fall into a period in viewers' lifespan that cognitive psychology knows as the *reminiscence bump* [25]. Empirical findings indicate that memories made during this period of early adulthood (i.e. between age 15 to 30) remain particularly accessible throughout people's lives. We expect this measure to maximize the possibility of viewers possessing accessible memories associated with our videos.

All subjects that we recruited gave their informed consent before entering the study itself, both regarding the tasks involved and the usage of their data. For a detailed overview about the demographics of the participating crowd-workers see the relevant section in *Table 4.1*. Because of the small number of participants in our dataset that are not from either the United States of America or The Republic of India (i.e. the *OTHER*-category in *Table 4.1*), we exclude all data belonging to these participants from the analysis reported throughout this article. This filtering reduces the total amount of individuals for which data is available to 288 unique viewers.

4.3.3. PROCEDURE AND APPARATUS

We developed an online application which crowd-workers could access through their web browser. It guided them through the entire data collection study, presented them with survey elements to acquire self-reports, and created audiovisual recordings using their webcams during their exposure to the videos.

After providing their informed consent for participation in the study, crowd-workers filled out an initial survey with background information about themselves (*Viewer-specific Self-Reports*). Then each person was provided with a random selection of 7 music videos from within our pool of 42 candidates. Acquisition of data about viewers' responses to each video had the following structure: first, they were watching the video, during which we recorded their upper-body behaviour with their device's webcam. Immediately after the video finished playing, the application collected self-reports about their experience (*Response-specific Self-reports*). This required participants to start by rating the emotional impact that the video had on them (its *Induced Emotion*). Then they were requested to report whether the video had evoked memories in them (i.e. whether any *Recollection* had occurred at all during exposure). If this was the case, we required participants to fill in a detailed survey to describe each of these memories in more detail, including the feelings that they associated with the memory (i.e. *Memory-Associated Affect*). After responding to all selected videos in this way, they had completed the data collection, and we debriefed them.

4.3.4. SELF-REPORT MEASURES

Here we provide a detailed description of all self-report measures that we have collected. Descriptive statistics can be found in *Table 4.1* for viewer-specific, and in *Table 4.2* for response-specific measures.

Table 4.1: Overview of Viewer-Specific Measures

DEMOGRAPHICS						
	USA		**India**		**Other**	
	N_f	N_m	N_f	N_m	N_f	N_m
Gender	125	115	11	37	2	10
	USA		**India**		**Other**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Age	33.57	6.01	30.54	4.91	32.83	4.91
PERSONALITY						
	USA		**India**		**Other**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Honesty	2.78	0.72	1.95	0.55	2.56	0.55
Emotional.	1.95	0.8	1.85	0.62	2.04	0.62
Extraver.	2.54	0.78	2.32	0.67	2.56	0.67
Agreeabl.	2.08	0.66	2.05	0.68	1.75	0.68
Conscien.	2.68	0.71	2.36	0.62	2.46	0.62
Openness	2.78	0.7	2.69	0.53	2.67	0.53
MOOD						
	USA		**India**		**Other**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Pleasure	0.42	0.4	0.43	0.44	0.36,	0.44
Arousal	-0.14	0.77	0.05	0.82	-0.38,	0.82
Dominance	0.34	0.49	0.51	0.47	0.46	0.47

Measures taken once per viewer: $N = 300$

VIEWER-SPECIFIC MEASURES

Demographics: In previous studies, demographic information significantly accounted for variation in viewers' emotional responses (e.g. age [5]). We capture self-reports of the following basic features: participants' age in years, their gender, and their nationality.

Personality: We collected data about our participants' personality traits in terms of the HEXACO scheme. It is a framework that aims to account for a wide variety of individual differences across peoples' behaviours by differentiating between them with a set of stable personality traits. In the HEXACO scheme these traits are defined by six orthogonal dimensions: (1) Honesty-Humility (H), (2) Emotionality (E), (3) eXtraversion (X), (4) Agreeableness (A), (5) Conscientiousness (C), and (6) Openness to experience (O). In our study, we assessed viewers' HEXACO scores using the Brief HEXACO Inventory (BHI) [26]. Because its design is specifically aiming for brevity (it consists of only 24-items) without sacrificing validity, it is well suited for crowd-sourcing scenarios. Scores are continuous values in the range of $[1-5]$.

Table 4.2: Overview of Response-specific Measures

	INDUCED EMOTION					
	USA		**India**		**Other**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Pleasure	0.18	0.52	0.35,	0.54	0.22,	0.51
Arousal	-0.17	0.78	0.32,	0.73	-0.23,	0.78
Dominance	0.12	0.58	0.3,	0.64	0.23,	0.58
	MEMORY ASSOCIATED-AFFECT					
	USA		**India**		**Other**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Pleasure	0.31	0.53	0.48,	0.55	0.35,	0.61
Arousal	0.01	0.79	0.38,	0.71	0.07	0.76
Dominance	0.28	0.57	0.42,	0.62	0.36,	0.57

Measures taken once per response: $N = 2098$

Mood: Before we exposed participants to any videos, they provided affective ratings for their overall mood for the same day. Findings show that mood has a significant influence on the emotions that videos induce in viewers [6]. Mood ratings in our corpus take the form of pleasure-, arousal- and dominance-ratings on a continuous scale in the interval of $[-1,+1]$. We obtain them from participants with the *AffectButton* instrument – an interactive widget displaying an iconic facial expression which changes in response to mouse or touch interaction. Users can select the facial expression best fitting the affective judgment that they need to provide (see [27] for a detailed description and a validation study). We opted for this instrument because it allowed viewers without knowledge of the underlying psychological framework to provide quick and implicit PAD-ratings through choosing a face.

RESPONSE-SPECIFIC MEASURES

Induced Emotions: We also capture viewers' ratings for their emotional response to a video with the AffectButton instrument. Consequently, they are on a continuous scale for pleasure, arousal and dominance bounded by the $[-1,+1]$ interval. See *Table 4.2* for relevant details about their distribution.

Memory-Associated Affect: Likewise, we measured the feelings that viewers associate with recollected memory content with the AffectButton instrument. Our dataset contains a total of 944 instances in which participants' recollected at least one memory in response to a music video. In principle, participants could report as many memories as were triggered in them by each video. However, only about 6% of recollections involved more than 2 memories (a total of 53 instances). In these cases, we decided to choose the most intense memory involved to represent their collective emotional meaning. This choice is motivated by empirical findings from psychology research, highlighting the dominance of

emotionally intense parts of an event in retrospective summary judgments of emotional meaning (for example, [28]). To make this selection, we rank all memories involved in a given multi-memory recollection based on the following measure for the affective intensity (I) of their associated affect: $I = \sqrt{\left(p^2 + ((a+1)/2)^2 + d^2\right)}$, where p, a, and d are the pleasure, arousal and dominant components of a particular rating. Note that this formula interprets negative arousal values as low-intensity experiences. This choice is motivated by the layout of the AffectButton instrument, which maps maximum negative arousal to neutral faces [27].

After calculation, we retain only the highest-scoring memory for further modelling activities. When we use *Memory-Associated Affect* in the remainder of this manuscript (including *Table 4.2*), we refer to the ratings selected in this way.

4.4. THE IMPACT OF MEMORY PROCESSES ON VIDEO-INDUCED EMOTIONS

In this section we present two empirical investigations to quantify the impact of personal memories on viewers' emotional responses during video-viewing scenarios in-the-wild.

4.4.1. EXP. 1: VIDEO-INDUCED EMOTIONS DIFFER WHEN PERSONAL MEMORIES ARE RECOLLECTED

In this first experiment, we investigate whether there are differences in the emotional responses of viewers' to a video (i.e. their Induced Pleasure, Induced Arousal, or Induced Dominance), depending on whether it made them recollect personal memories or not. Moreover, we explore the degree to which these differences depend on the identity of eliciting videos.

Method and Approach: We use linear mixed regression models in our analysis to account for the repeated measures of responses from the same viewers in our data collection. A separate model was fitted for each affective dimension of viewers' induced emotions: *Induced Pleasure*, *Induced Arousal*, and *Induced Dominance*. The fixed effects included in these models are the *identity of the video (VID)* shown to viewers, the *Occurrence of Recollection (REC)*, as well as their two-way interaction term. *REC* is a binary variable denoting whether a viewer's response involved the recollection of a video or not. *VID*, on the other hand, is a factor with 42 levels, each denoting the identity of the particular music video that we showed to participants. Additionally, we specify viewers' identity as a random effect in the models, thereby accounting for the dependence among their responses due to repeated measures.

Results: We conducted an analysis of variance for the fixed effects in the separate models (*Table 4.3*). To account for the multiple comparisons between the same set of dependent variables with each of the three independent variables, we have applied Bonferroni corrections to all statistical tests. Results show that the occurrence of recollection has a significant effect on video-induced emotion across all dimensions. This finding indicates that there is a difference in video-induced emotions when viewers' responses also involved the recollection of personal memories. To investigate the direction of these differences,

we compared the means of responses involving recollections to those without them. Results indicate that when recollections are present ratings for induced emotions are higher across all affective dimensions (induced pleasure: $M\Delta = +.16$, $t(2043.7) = 7.05$, $p =< .001$; induced arousal: $M\Delta = +.17$, $t(2033) = 5.09$, $p =< .001$; induced dominance: $M\Delta = +.19$, $t(2058.2) = 7.46$, $p =< .001$). Moreover, the absence of a significant interaction-effect reveals that the magnitudes of these differences are comparable across videos.

Finally, we investigated the increase in explained variance contributed by each of the fixed effects in the model to gain insights into their relative explanatory power. For this purpose we use a measure for the explained variance of the fixed effects in linear mixed models – Marginal R^2 (R^2_m) [29]. Consequently, the measure of ΔR^2_m in *Table 4.3* captures the additional variance explained by a model that includes the predictors for a given fixed effect, compared to one that does not. A comparison of this metric across models for the different affective dimensions reveals that video identity is the effect explaining the highest amount of unique variance ($Avg\Delta R^2_m = .13$). While the effects of occurring recollections are significant, they explain only a rather small amount of unique variation ($Avg\Delta R^2_m = .017$).

4.4.2. Exp. 2: Memory-Associated Affect Predicts Video-Induced Emotions

Psychological findings indicate that the affect associated with the memory content triggered by media stimuli correlates with the emotions that these stimuli induce in them [17]. Here, we explore the existence of this relationship for those instances in our dataset, where viewers' have recollected memory content in response to music videos. In particular, we assess whether their feelings towards the memory content evoked by a particular video (i.e. *Memory-associated Affect*) are indicative of their emotional response to it.

Method and Approach: Like in the previous experiment we fit three separate linear mixed regression models for our analysis, each targeting one of the affective dimensions of viewers induced emotions (i.e. *Induced Pleasure*, *Induced Arousal*, and *Induced Dominance*. All models include fixed effects for 1. video identity VID (again a factor with 42 levels), and 2. the memory-associated affect MA, which consists of ratings for pleasure, arousal and dominance as predictors. Additionally, we include all two-way interactions between the predictors of the specified VID and MA-effects. All affective ratings for induced emotion and memory-associated affect are continuous numerical variables and constrained to the interval $[-1, 1]$. Before introducing them to the model, they are standardized by subtracting their mean and dividing by their standard deviation. We include participants' identity as a random effect to account for the repeated-measures design of our data collection.

Results: We conduct an analysis of variance for the fixed effects in the specified models, the results of which we present in *Table 4.4*). We applied Bonferroni correction when testing for significance to account for multiple comparisons between dependent and independent variables across models. Significant effects exist both for memory-associated affect and video identity across all models for induced emotion. Moreover, we find significant two-way interactions between video identity and memory-associated affect.

Table 4.3: Exp. 1 – Effect of Occurring Recollections on Video-Induced Emotion.

		Induced Pleasure				Induced Arousal				Induced Dominance				
Effect	df_n	df_d	F	p	ΔR^2_m	df_d	F	p	ΔR^2_m	df_d	F	p	ΔR^2_m	$Avg\Delta R^2_m$
VID	41	1937.48	10.27	<.001*	.16	1864.78	6.43	<.001*	.1	1878.43	8.74	<.001*	.13	.13
REC	1	1683.29	47.1	<.001*	.02	1977.8	30.96	<.001*	.01	1942.93	38.48	<.001*	.02	.017
$VID * REC$	41	1957.52	1.29	.106	.02	1880.93	1.28	.11	.02	1895.85	1.28	.115	.02	.02

* Result below Bonferroni corrected critical value for significance: $\alpha_{adj} = .017$

Table 4.4: Exp. 2 – Effect of Memory-associated Affect on Video-Induced Emotion.

		Induced Pleasure				Induced Arousal				Induced Dominance				
Effect	df_n	df_d	F	p	ΔR^2_m	df_d	F	p	ΔR^2_m	df_d	F	p	ΔR^2_m	$Avg\Delta R^2_m$
VID	41	877.07	2.64	<.001*	.05	865.28	1.36	.067	.03	885.22	1.86	.001*	.04	.04
MA	3	889.39	228.92	<.001*	.34	895.33	141.62	<.001*	.29	878.35	161.02	<.001*	.29	.306
$VID * MA$	123	738.16	1.4	.005*	.08	733.37	1.34	.014*	.11	738.99	1.63	<.001*	.11	.1

* Result below Bonferroni corrected critical value for significance: $\alpha_{adj} = .017$

This finding indicates video-specific variations in the strength of the relationship between the emotions induced in viewers and their feelings towards recollected memories. Finally, we compare the changes in uniquely explained variance for each of the fixed effects in the specified models. This reveals that memory-associated affect makes the strongest contribution ($Avg\Delta R^2_m = .306$), going beyond the unique share of video identity ($Avg\Delta R^2_m = .04$). Moreover, the interactions between these effects explain a relatively large share of additional variance in viewers' responses ($Avg\Delta R^2_m = .1$).

4.4.3. DISCUSSION

Our first experiment demonstrates that peoples' experience of a video significantly differs depending on whether it triggers personal memories or not. This finding points to a recollection-specific bias, causing videos that trigger memories to display induced emotions with heightened levels of induced pleasure, arousal and dominance. This finding confirms the results of existing psychological research involving music-evoked recollections (e.g. [10], and shows that such biases are also present for video material the unconstrained scenarios captured by our dataset. More generally, our results highlight the existence of systematic influences on the emotions induced by videos that one cannot feasibly attribute to their content alone but instead result from effects in the situation under which viewing takes place. Naturally, technologies for emotion prediction that solely rely on the analysis of the audiovisual signals comprising this content – as is the dominant approach in VACA – cannot account for this kind of influence.

In our second experiment, we have discovered a strong relationship between the affect that a viewer associates with a video-triggered memory and the triggering video's emotional emotional impact on him/her. This effect explains an amount of variation that goes significantly above and beyond what one can reasonably attribute to video-specific influences alone, further underlining the importance of accounting for memories when modelling individuals' experiences. Moreover, this relationship varies in intensity across different videos.

A possible explanation for this is that the memories evoked by some videos are more engaging as a target for viewers than their audiovisual content. Exposure to these videos could create conditions where viewers' attention is more likely to drift inwards, thereby increasing memories' emotional impact. Such mind-wandering phenomena can occur across different forms of media consumption, e.g. reading, and its emergence depends on individuals' availability of attentional resources [30]. Being able to identify videos that are less impacted by memory-associated affect automatically could be a valuable effort for computational research because it is likely that responses to these are more firmly grounded in a video's audiovisual content. Such a grounding matches the stimulus-centric modelling assumption of direct VACA, and consequently applying it to these videos might improve results.

Together, both studies demonstrate the scope and depth of the role that personal memories play in viewers' experiences. Moreover, they point towards the potential that information about individuals' recollection processes could hold as context for predictions of video-induced emotions in technological systems.

4.5. USING MEMORIES TO PERSONALIZE PREDICTIONS OF INDUCED EMOTIONS

We found that memory-associated affect strongly correlates with video-induced emotion. Therefore, in this section, we explore to what extent the occurrence of recollections and memory-associated affect influence the accuracy of predicting video-induced emotion. We assess the relative contribution of these memory-related features for personalizing predictions compared to information about viewers' demographics, their personality traits, and their overall mood at the time of exposure to the video.

4.5.1. EXP. 3: OCCURRENCE OF RECOLLECTIONS

In this experiment, we explore the increase in predictive power provided solely by information about the occurrence of recollections in response to videos.

Method and Approach: We constructed a linear mixed regression model for each affective dimension of the emotions induced in viewers: *Induced Pleasure*, *Induced Arousal*, and *Induced Dominance*. The fixed effect Occurrence of Recollections (REC) is a factor with two levels, capturing the presence or absence of personal memories as part of the response to a video. A fixed-effect for the identity of the video (VID), consisting of a factor with 42 levels, captures information that can be feasibly provided by the video content itself. Additionally, we define the following fixed effects for various characteristics collected as viewer-specific measures (see Table 4.1):

1. Demographics (DEM) includes a continuous predictor for viewers' age, one 2-level factor representing viewers' nationality (USA/India), and another 2-level factor for their gender (male/ female).

2. Personality (PER) comprises five continuous predictors, one for each of the HEXACO personality traits

3. Mood ($MOOD$) includes three predictors, one each for viewers' self-reported pleasure, arousal, and dominance

We standardized the continuous predictors in all of the specified fixed effects, as well as target variables, by subtracting their respective mean and dividing by their standard deviation before introducing them into the regression model. Finally, all models include participants' identity as a random effect to account for the repeated measures of responses.

Results: The results for an analysis of variance of the fixed effects specified in our regression models are presented in *Table 4.5*. All tests for significance are Bonferroni corrected to account for multiple comparisons between the predictors and the target variables across the different models. Our findings indicate that information about recollection occurrences contributes to the accuracy of models above and beyond the other included sources. However, the amount of unique variation explained by it is rather small ($Avg\Delta R^2_m = .016$). A look at the remaining effects in the models reveals that

Table 4.5: Exp. 3 – Comparisons including the Effect of Occurring Recollections on Video-Induced Emotion

		Induced Pleasure				Induced Arousal				Induced Dominance				
Effect	df_n	df_d	F	p	ΔR^2_m	df_d	F	p	ΔR^2_m	df_d	F	p	ΔR^2_m	$Avg\Delta R^2_m$
VID	41	1915.96	9.8	<.001*	.159	1852	6.1	<.001*	.095	1852.99	8.34	<.001*	.126	.126
DEM	3	277.28	8.88	<.001*	.012	277.06	12.39	<.001*	.026	277.07	3.18	.025	.007	.012
PERS	6	275.43	1.38	.222	.004	275.48	1.39	.22	.007	275.48	1.85	.09	.008	.006
MOOD	3	275.2	4.6	.004*	.007	275.29	8.01	<.001*	.021	275.29	6.21	<.001*	.015	.014
REC	1	1629.06	45.58	<.001*	.018	1889.19	29.8	<.001*	.013	1886.63	38.23	<.001*	.017	.016

* Result below Bonferroni corrected critical value for significance: $\alpha_{adj} = .017$

Table 4.6: Exp. 4 – Comparisons including the Effect of Memory-associated Affect on Video-Induced Emotion

		Induced Pleasure				Induced Arousal				Induced Dominance				
Effect	df_n	df_d	F	p	ΔR^2_m	df_d	F	p	ΔR^2_m	df_d	F	p	ΔR^2_m	$Avg\Delta R^2_m$
VID	41	1921.56	6.67	<.001*	.092	1853.22	4.75	<.001*	.064	1848.95	5.47	<.001*	.071	.075
DEM	3	277.51	9.18	<.001*	.011	277.3	11.16	<.001*	.024	277.28	3.07	.028	.006	.013
PERS	6	275.67	1.91	.08	.005	275.7	0.8	.568	.004	275.69	2.28	.037	.008	.005
MOOD	3	277.3	2.86	.037	.004	276.82	6.01	.001*	.016	276.78	4.7	.003*	.012	.014
MEM	8	1933.19	49.13	<.001*	.134	1919.4	35.56	<.001*	.104	1916.07	42.19	<.001*	.115	.117

* Result below Bonferroni corrected critical value for significance: $\alpha_{adj} = .017$

video identity provides the biggest insights into viewers' responses across all affective dimensions ($Avg\Delta R^2_m = .126$). Note that the results for these effects are different from those in experiment 2 (see *Table 4.4*) because models for the current analysis include viewer-characteristics as additional predictors and do not contain interaction effects. Moreover, the analysis reveals that demographics explain a small amount of variation ($Avg\Delta R^2_m = .012$) in induced pleasure and arousal, but do not contribute to predictions of dominance. Mood consistently provides a degree of information about emotional responses comparable to that of occurring recollection ($Avg\Delta R^2_m = .014$) The average overall fit for the fixed effects specified in the models was modest ($AvgR^2_m = .2$).

4.5.2. EXP. 4: MEMORY-ASSOCIATED AFFECT

In this experiment, we assess the added benefits for personalizing predictions gained by information about the affect associated with a recollected memory.

Method and Approach: We specify separate linear mixed-effects regression models for each of the affective dimensions of viewers' induced emotions. Each includes the fixed effects specified in Exp. 4 for viewers' demographics (DEM), their personality traits ($PERS$), and their mood ($MOOD$), as well as the effect of video identity (VID). In addition to these, all models include a fixed effect for memory-associated affect (MEM), which is a factor with nine levels. It denotes whether a response either 1. involves no recollection, or 2. the octant of the affective rating associated with the recollected memory in the three-dimensional PAD-space (e.g. a memory associated with positive values for pleasure, arousal, and dominance would be assigned to the octant $P_+A_+D_+$). All models include a random effect for viewers' identity to account for the repeated measures in our design. This coding allows models to predict both responses involving memories, and those that do not.

Results: *Table 4.6* displays the results of an analysis of variance for the fixed effects in the models specified to investigate the predictive power of Memory-associated Affect (Bonferroni corrected). It shows that information about viewers' memories accounts for the greatest share of explained variance across models (MEM: $Avg\Delta R^2_m = .117$), outperforming that of stimulus identity (VID: $Avg\Delta R^2_m = .075$). Again, demographics offered some information about viewers' pleasure and arousal, but not dominance. In contrast to the model for the occurrence of recollections, the contributions of mood are no longer significant for explaining pleasure. Just as before, the effect of viewers' personality is not significant. With an average of $AvgR^2_m = .3$ for the fixed effects across models, they fit the data better than those specified in experiment 3. Nevertheless, they failed to account for the majority of the variation in viewers' responses.

4.5.3. DISCUSSION

While the occurrence of recollections offers only minor contributions to predictions about viewers' emotional responses to videos, memory-associated affect emerged as the strongest predictor. Moreover, these contributions go above and beyond those offered by viewer-characteristics, or the identity of the eliciting music video.

Providing intelligent applications with information about viewers' recollection processes at prediction time in the dynamic fashion assumed in our analyses is a challenging problem because it requires technology that is both able to meaningfully estimate 1. when recollections do occur and 2. what affect viewers' associate with the evoked memories. However, despite these substantial obstacles, recollection and memory-associated affect are the most important source of information for predicting video-induced emotions, above and beyond the video itself, personality, mood, and demographics.

Our research shows that not addressing the influence of memories will always limit the accuracy of automatic predictions of video-induced affect. In and of itself, this is not a problem, but something to be aware of when developing such systems.

Finally, the overall modest fit across our models in both analyses (occurrence of recollections: $AvgR^2_m = .2$; memory-associated affect: $AvgR^2_m = .3$) points towards significant room for improvement by incorporating additional viewer- and situation-specific features.

Empirical research from cognitive psychology has described a wide variety of contextual influences on acting on human cognitive-affective processing that could be explored in computational models of video-induced emotions, e.g. the presence of enduring goals and values [4]. An essential step in this direction is the careful development of datasets of emotional responses that systematically capture relevant contextual attributes [6, 31]. Importantly, the overall limited insights offered by static viewer-characteristics in our analyses underline that such efforts should focus increasingly on dynamic attributes of viewers, their cognitions, and the situations in which they take place.

4.6. Limitations

There are notable limitations to how our empirical findings generalize to other types of media material or a different viewership. First, our data collection involves only responses to a particular format of media content, i.e. music videos. It is plausible that the connection between personal memories and emotional impact is less profound for other content formats. For example, in feature films, empathy with the portrayed characters in the narrative is a critical aspect [32] that could overshadow the influence of any personal memories. It is also important to point out that we have purposefully selected both participants and content to increase the chance for recollection. While this should not have an impact on the validity of our findings regarding the effects of memories, our data may display an inflated rate of their occurrence. However, a realistic understanding of the conditions under which videos trigger memories in members of the general population requires a more diverse content-participant mixture.

4.7. Towards Addressing Memory-Influences in Automatic Predictions

Attempting to provide applications with information about viewers' recollections offers numerous opportunities for empirical and technological exploration. For example, there exists no direct computational research modelling the evocative properties of videos or of the situations in which people view them. Similar to existing work estimating the

likelihood of media content to be remembered (e.g. [33]), modelling videos' capacity to trigger memories could be explored based on their audiovisual features. Additionally, Given that involuntary memory recollection has been connected to attentional drifting, work on predicting such mind-wandering from multimodal data can serve as a starting point for modelling viewers' pensiveness in a situation. In this setting, researchers have successfully used measures of an individual's physiology or overt behaviour to detect when attention is turning inwards and away from video content to other thoughts [34].

Likewise, existing work from ubiquitous computing, sentiment analysis, and cognitive modelling can form the basis for predicting the affect viewers' associate with their memories (see [9] for a detailed discussion). Such efforts could centre around personal data that has been either collected implicitly by rich ubiquitous sensing (e.g. through lifelogging [35]), or provided explicitly as social media content (e.g. comments in response to media [36], or as entries in smart journals [37]).

4.8. SUMMARY AND CONCLUSION

In this article, we have presented two contributions relevant to predictions of video-induced emotions: (1) two empirical investigations exploring the effects that being reminded of personal memories by a video has on induced emotions, and (2) two additional experiments in which we explore the relative value of access to features describing viewers' recollection processes for understanding variations in their emotional responses.

The findings of our first set of experiments show that the presence of associated personal memories coincides with a stronger emotional impact on them, independently of the video that is being viewed. This indicates that recollections are a ubiquitous influence on viewers' subjective experience of video material. Moreover, when memories are triggered, induced emotions are often similar to the affect that viewers associate with what has been remembered. However, the degree of this similarity varies across videos, showing that viewers' experience of some videos is more strongly influenced by their memories than that of others. As a consequence, one goal for technological research should be to automatically detect the importance of memories for experiencing a particular video.

In our second set of experiments, we found that both the occurrence of and affect-associated with personal memories explain variation in viewers' emotional responses to videos above and beyond the video itself and relevant viewer-characteristics. These results indicate that providing this information to computational models holds significant potential for predictions of subjective viewing experiences. Moreover, the negligible contribution of static viewer-characteristics to predictions of induced emotions (e.g. personality traits), highlights the necessity of access to such highly situation-specific information. Consequently, without accounting for dynamic influences like personal memories in computational models, accurate predictions of video-induced emotions in real-life applications will remain out of reach. This is a challenging endeavor, but we have outlined several lines of existing technological research that can form a starting point for exploring automatic predictions of when memories occur, and how they impact a viewers' experience. As such, progress seems difficult, but possible.

REFERENCES

[1] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, *Affective video content analysis: A multidisciplinary insight,* IEEE Transactions on Affective Computing **9**, 396 (2018).

[2] M. Soleymani, M. Pantic, and T. Pun, *Multimodal Emotion Recognition in Response to Videos,* IEEE Transactions on Affective Computing **3**, 211 (2012).

[3] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, *Multimedia content analysis for emotional characterization of music video clips,* EURASIP Journal on Image and Video Processing **2013**, 26 (2013).

[4] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, *Appraisal Theories of Emotion: State of the Art and Future Development,* Emotion Review **5**, 119 (2013).

[5] 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 **13**, 212 (2012).

[6] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, *Corpus Development for Affective Video Indexing,* IEEE Transactions on Multimedia **16**, 1075 (2014), arXiv:1211.5492 .

[7] M. J. Scott, S. C. Guntuku, W. Lin, and G. Ghinea, *Do Personality and Culture Influence Perceived Video Quality and Enjoyment?* IEEE Transactions on Multimedia **18**, 1796 (2016).

[8] Y. Zhu, I. Heynderickx, and J. A. Redi, *Understanding the role of social context and user factors in video Quality of Experience,* Computers in Human Behavior **49**, 412 (2015).

[9] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, *Artificial Empathic Memory,* in *Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions - EE-USAD'18* (ACM Press, New York, New York, USA, 2018) pp. 1–8.

[10] P. Janata, S. T. Tomic, and S. K. Rakowski, *Characterisation of music-evoked autobiographical memories,* Memory **15**, 845 (2007).

[11] D. Berntsen, S. R. Staugaard, and L. M. T. Sørensen, *Why am I remembering this now? Predicting the occurrence of involuntary (spontaneous) episodic memories.* Journal of Experimental Psychology: General **142**, 426 (2013).

[12] D. G. McDonald, M. A. Sarge, S.-F. Lin, J. G. Collier, and B. Potocki, *A Role for the Self: Media Content as Triggers for Involuntary Autobiographical Memories,* Communication Research **42**, 3 (2015).

[13] A. M. Belfi, B. Karlan, and D. Tranel, *Music evokes vivid autobiographical memories,* Memory **24**, 979 (2016).

[14] J. H. Mace, *Involuntary autobiographical memories are highly dependent on abstract cuing: the Proustian view is incorrect,* Applied Cognitive Psychology **18**, 893 (2004).

[15] D. van Gennip, E. van den Hoven, and P. Markopoulos, *Things That Make Us Reminisce,* in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, Vol. 1 (ACM Press, New York, New York, USA, 2015) pp. 3443–3452.

[16] C. Mills and S. D'Mello, *On the Validity of the Autobiographical Emotional Memory Task for Emotion Induction,* PLoS ONE **9**, e95837 (2014).

[17] H. Baumgartner, *Remembrance of things past: Music, autobiographical memory, and emotion,* Advances in Consumer Research **19**, 613 (1992), arXiv:9301106587 .

[18] S. Wang and Q. Ji, *Video Affective Content Analysis: A Survey of State-of-the-Art Methods,* IEEE Transactions on Affective Computing **6**, 410 (2015).

[19] D. McDuff and M. Soleymani, *Large-scale Affective Content Analysis: Combining Media Content Features and Facial Reactions,* in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (IEEE, 2017) pp. 339–345.

[20] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *DEAP: A Database for Emotion Analysis ;Using Physiological Signals,* IEEE Transactions on Affective Computing **3**, 18 (2012).

[21] 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 *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2015) pp. 236–242.

[22] P. Ekman, *Basic emotions,* (1999).

[23] J. A. Russell, *A circumplex model of affect.* Journal of Personality and Social Psychology **39**, 1161 (1980).

[24] A. Mehrabian, *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament,* Current Psychology **14**, 261 (1996).

[25] M. A. Conway and S. Haque, *Overshadowing the Reminiscence Bump: Memories of a Struggle for Independence,* Journal of Adult Development **6**, 35 (1999).

[26] R. E. de Vries, *The 24-item Brief HEXACO Inventory (BHI),* Journal of Research in Personality **47**, 871 (2013).

[27] J. Broekens and W.-P. Brinkman, *AffectButton: A method for reliable and valid affective self-report,* International Journal of Human-Computer Studies **71**, 641 (2013).

[28] A. M. DO, A. V. RUPERT, and G. WOLFORD, *Evaluations of pleasurable experiences: The peak-end rule,* Psychonomic Bulletin & Review **15**, 96 (2008).

[29] S. Nakagawa and H. Schielzeth, *A general and simple method for obtaining R 2 from generalized linear mixed-effects models,* Methods in Ecology and Evolution **4**, 133 (2013).

[30] S. Feng, S. D'Mello, and A. C. Graesser, *Mind wandering while reading easy and difficult texts,* Psychonomic Bulletin & Review **20**, 586 (2013).

[31] B. Dudzik, M.-P. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong, *Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases,* in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 206–212.

[32] J.-J. Igartua, *Identification with characters and narrative persuasion through fictional feature films,* Communications **35**, 347 (2010).

[33] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, *Understanding and Predicting Image Memorability at a Large Scale,* in *2015 IEEE International Conference on Computer Vision (ICCV)*, Vol. 2015 Inter (IEEE, 2015) pp. 2390–2398.

[34] A. Stewart, N. Bosch, H. Chen, P. Donnelly, and S. D'Mello, *Face Forward: Detecting Mind Wandering from Video During Narrative Film Comprehension,* in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10331 LNAI (Springer, Cham, 2017) pp. 359–370.

[35] C. Gurrin, A. F. Smeaton, and A. R. Doherty, *LifeLogging: Personal Big Data,* Foundations and Trends® in Information Retrieval **8**, 1 (2014).

[36] D. Cosley, V. Schwanda, S. T. Peesapaty, J. Schultz, and J. Baxter, *Experiences with a publicly deployed tool for reminiscing,* in *Proc. First Int'l Workshop on Reminiscence Systems* (2009) pp. 31–36.

[37] C. Elsden, A. C. Durrant, and D. S. Kirk, *It's Just My History Isn't It?* in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (ACM, New York, NY, USA, 2016) pp. 2819–2831.

5

Personal Memory Appraisal as Context for Video Affective Content Analysis

This chapter is based on: Dudzik, B., Broekens, J., Neerincx, M., & Hung, H. (2020). A Blast From the Past: Personalizing Predictions of Video-Induced Emotions using Personal Memories as Context. ArXiv. https://arxiv.org/abs//2008.12096

ABSTRACT

A key challenge in the accurate prediction of viewers' emotional responses to video stimuli in real-world applications is accounting for person- and situation-specific variation. An important contextual influence shaping individuals' subjective experience of a video is the personal memories that it triggers in them. Prior research has found that this memory influence explains more variation in video-induced emotions than other contextual variables commonly used for personalizing predictions, such as viewers' demographics or personality. In this chapter, we show 1. that automatic analysis of text describing their video-triggered memories can account for variation in viewers' emotional responses, and 2. that combining such an analysis with that of a video's audiovisual content enhances the accuracy of automatic predictions. We discuss the relevance of these findings for improving on state of the art approaches to automated affective video analysis in personalized contexts.

5.1. INTRODUCTION

The experience of specific feelings and emotional qualities is an essential driver for people to engage with media content, e.g. for entertainment or to regulate their mood [1]. For this reason, research on *Video Affective Content Analysis (VACA)* attempts to automatically predict how people emotionally respond to videos [2]. This has the potential to enable media applications to present video content that reflects the emotional preferences of their users [3], e.g. through facilitating emotion-based content retrieval and recommendation, or by identifying emotional highlights within clips. Existing VACA approaches typically base their predictions of emotional responses mostly on a video's audiovisual data [4].

In this chapter, we argue that many VACA-driven applications can benefit from emotion predictions that also incorporate viewer- and situation-specific information as context for accurate estimations of individual viewers' emotional responses. Considering context is essential, because of the inherently subjective nature of human emotional experience, which is shaped by a person's unique background and current situation [5]. As such, emotional responses to videos can be drastically different across viewers, and even the feelings they induce in the same person might change from one viewing to the next, depending on the viewing context [6]. Consequently, to achieve affective recommendations that meaningfully match a viewer's current desires, e.g., to see an "amusing" video, these will likely need some sensitivity for the conditions under which any potential video will create this experience specifically for him/her. Therefore, addressing variation in emotional responses by personalizing predictions is a vital step for progressing VACA research (see the reviews by Wang et al. [4] and Baveye et al. [2]).

Despite the potential benefits, existing research efforts still rarely explore incorporating situation- or person-specific context to personalize predictions. Reasons for this include that it is both difficult to *conceptualize* context (identifying essential influences to exploit in automated predictions), as well as to *operationalize* it (obtaining and incorporating relevant information in a technological system) [7]. Progress towards overcoming these challenges for context-sensitive emotion predictions requires systematic exploration in computational modeling activities, guided by research from the social sciences [8].

Here, we contribute to these efforts by exploring the potential to account for personal memories as context in automatic predictions of video-induced emotions. Findings from psychology indicate that audiovisual media are potent triggers for personal memories in observers [9, 10], and these can evoke strong emotions upon recollection [11]. Once memories have been recollected, evidence shows that they possess a strong ability to elicit emotional responses, hence their frequent use in emotion induction procedures [12]. Moreover, the affect associated with media-triggered memories – i.e., how one feels about what is being remembered – has a strong connection to its emotional impact [13, 14]. These findings indicate that by accessing the emotion associated with a triggered memory, we are likely to be able to obtain a close estimate of the emotion induced by the media stimulus itself. Moreover, they underline the potential that accounting for the emotional influence of personal memories holds for applications. They may enable technologies a more accurate overall reflection of individual viewers' emotional needs in affective video-retrieval tasks. However, they might also facilitate novel use-cases that target memory-related feelings in particular, such as recommending nostalgic media content [15].

One possible way to address memory-influences in automated prediction is through the analysis of text or speech data in which individuals explicitly describe their memories. Prior research has revealed that people frequently disclose memories from their personal lives to others [16, 17], likely because doing so is an essential element in social bonding and emotion regulation processes [18]. There is evidence that people share memories for similar reasons on social media [19], and that they readily describe memories triggered in them by social media content [15]. Moreover, predicting the affective state of authors of social media text [20], as well as text analysis of dialog content [21], are active areas of research. However, apart from the work of Nazareth et al. [22], we are not aware of specific efforts to analyze memories. Together, these findings indicate that it may be feasible to both (1) automatically extract emotional meaning from free-text descriptions of memories and (2) that such descriptions may be readily available for analysis by mining everyday life speech or exchanges on social media. This second property may also make memory descriptions a useful source of information in situations where other data may be unavailable due to invasiveness of the required sensors – e.g., when sensing viewers' facial expressions or physiological responses. Motivated by the potential of analyzing memories for supporting affective media applications, we present the following two contributions:

1. we demonstrate that it is possible to explain variance in viewers' emotional video reactions by automatically analyzing free-text self-reports of their triggered memories, and

2. we quantify the benefits for the accuracy of automatic predictions when combining both the analysis of videos' audiovisual content with that of viewers' self-reported memories

5.2. BACKGROUND AND RELATED WORK

5.2.1. VIDEO AFFECTIVE CONTENT ANALYSIS

With respect to the use of context, existing work can be categorized into two types: 1. *context-free VACA* and 2. *context-sensitive VACA*.

Context-Free VACA: Works belonging to this type simplify the task of emotion prediction by making the working assumption that every video has quasi-objective emotional content [4]. Traditionally, researchers define this content as the emotion that a stimulus results in for a *majority* of viewers (i.e. its *Expected Emotion* [2]). The goal of VACA technology then is to automatically provide a single label for a video representing its expected emotional content, while ignoring variation in emotional responses that a video might elicit. Existing technological approaches for this task primarily consist of machine learning models trained in a supervised fashion on human-annotated corpora [4]. The ground truth for expected emotions is formed by aggregating the individual emotional responses from multiple viewers for the same video (e.g., by taking the mean or mode across the distribution of their responses). These models can then be used to automatically label entire databases of videos with tags representing their emotional content. The whole process of prediction of an individual viewer's emotional response using context-free VACA consists of two primary stages, as shown in the graphical overview displayed in *Figure 5.1*. 1. using a pre-trained VACA model to automatically label any video in a database of interest (with its expected emotion), and then 2. relying on the video label to be a plausible approximation for the specific emotion that any individual viewer experiences (i.e., his/her *Induced Emotion*).

There are two groups of technological approaches to automatically label videos in this way, differing in the information that they rely on as input for their predictions [4]. The first, and traditionally most widespread approach is *Direct VACA* (see region *A* in 5.1), which exclusively uses features derived from the audio and visual tracks of the video stimulus itself as the basis for predictions. The second is *Indirect VACA*, which is looking into automated approaches for analyzing the spontaneous behavior displayed by viewers to label videos without having to ask them for a rating. In essence, this approach uses measures of physiological responses or human behavior (e.g. [23, 24]) from a sample audience to predict the expected emotion of a population for a video (see region *B* in *Figure 5.1*). As such, this endeavor is closely connected to the broader research on emotion recognition from human physiology and behavior in affective computing (see D'mello and Kory [25] for a recent overview). Unfortunately, during the writing of this chapter, we found that the conceptual differences between the two research efforts are not necessarily clearly defined. Therefore, in this chapter, we define *emotion recognition* to be approached using only measures of an individual's behavior or physiology to predict his/her specific emotional response to a video (e.g., as in Shukla et al. [26]). In contrast to this, Indirect VACA methods (also known as *Implicit Tagging* [27]) collect behavioral or physiological data from multiple viewers in response to videos and use an aggregate of these measurements as input to label videos with their *expected* emotion automatically.

Context-free approaches intentionally blend out issues of subjectivity and variation. In essence, they rely entirely on the *expected* emotion-labels to be reasonable approximations for the emotional responses of an individual viewer to the video, independent of who he or she is, or in what situation they are watching the video. Because the prediction target is a video-specific aggregate, context-free affective content can only be valid if stimuli display only a small amount of within-video variation in responses. Creators of corpora for VACA modeling have typically enforced this by selecting only stimulus videos for which a low degree of induced emotional variation was already observed (e.g., [23, 28, 29]). Nev-

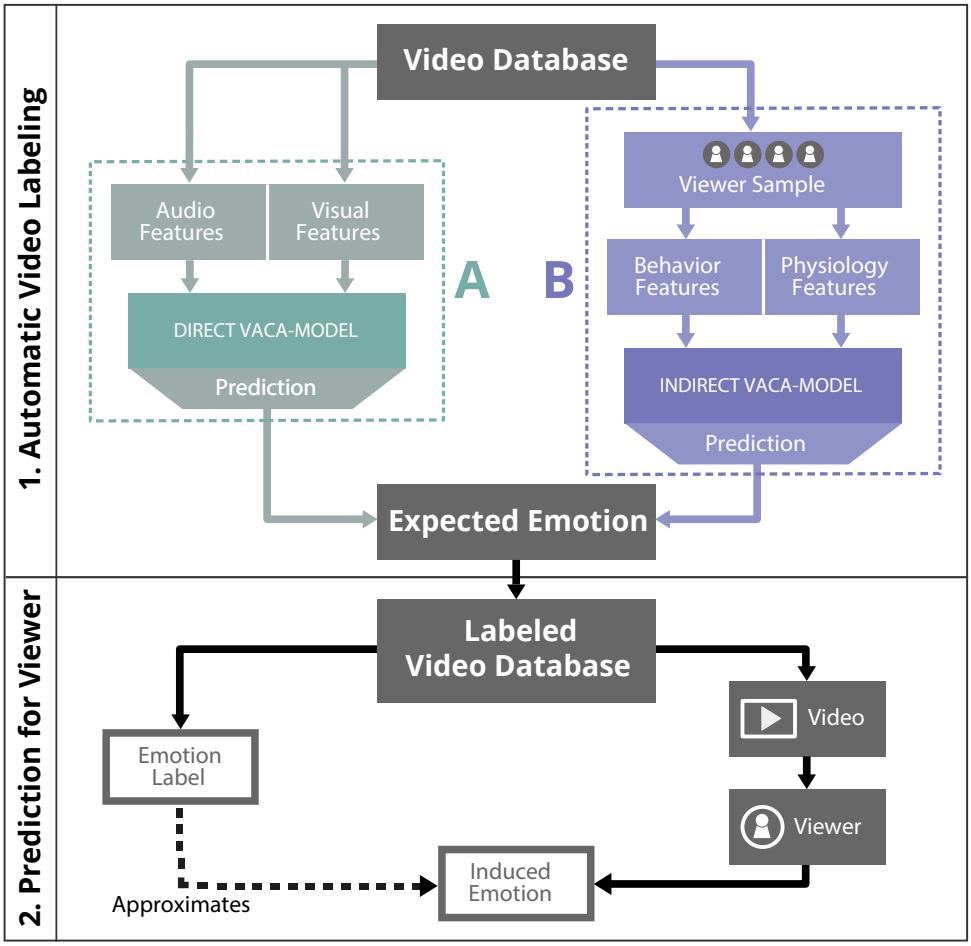


Figure 5.1: Overview about the Context-Free VACA approach – A: Direct VACA; B: Indirect VACA

ertheless, in naturalistic viewing conditions, such strongly homogeneous responses are unrealistic [6]. We argue that variation in emotional impact is the norm rather than the exception. Consequently, VACA research needs to expand its notion of affective content from being quasi-objective to inherently subjective and context-dependent if it wants to make predictions that meaningfully reflect individual's media experiences in the real world.

Context-sensitive VACA: We consider approaches as context-sensitive VACA when they 1. attempt to predict an individual viewer's affective responses to a video (i.e. the *Induced Emotion* [2]), and 2. rely on both the analysis of a video's audiovisual content and the context for this task. Existing works can further be distinguished according to the type of context that they use:

- *Viewer-specific context* refers to properties and traits of individual viewers aimed at accounting for variation in video-induced emotions between different individuals (e.g. demographics, personality), while
- *Situation-specific context* denotes information that is temporarily relevant for predicting a viewer's emotional responses. It covers influences that stretch over multiple videos viewed by the same person in succession (e.g. the type of social setting in which viewing takes place), or that may be specific to only a single instance of viewing (such as a triggered memory).

See *Figure 5.2* for a schematic overview of context-sensitive VACA.

According to this coarse categorization, initial efforts have explored the impact of person-specific context, capturing their personality and cultural background [30, 31]. Other research has touched on situational properties, such as mood or the time of day at which someone watches a video [6]. It has also explored the impact of viewing in a group compared to being alone [32]). Finally, measurements of an individual viewer's behavior or physiology while watching a video can be considered fine-grained information about their current situation and is explored extensively in existing computational approaches (e.g., [23, 33]).

5.2.2. REPRESENTING VIDEO-INDUCED EMOTIONS

Researchers in psychology do not universally agree upon a single system to formally categorize the various states and experiences that make up human emotion. Instead, there exist competing schemes and taxonomies, many of which have been used to represent video-induced emotion [2]. These fall broadly into two distinct groups: categorical and dimensional schemes. Categorical schemes describe human emotion in terms of a set of discrete emotional states, such as happiness or anger. In contrast to this, dimensional frameworks describe subjective emotional experiences as points in a continuous, multidimensional space. Each dimension in the scheme intends to capture a vital characteristic to differentiate emotional experiences from one another.

Computational work on video-induced emotions has traditionally favored categorical representation schemes [4], as does the field of automated emotion recognition in a broader sense [25]. However, psychological research has substantially criticized the

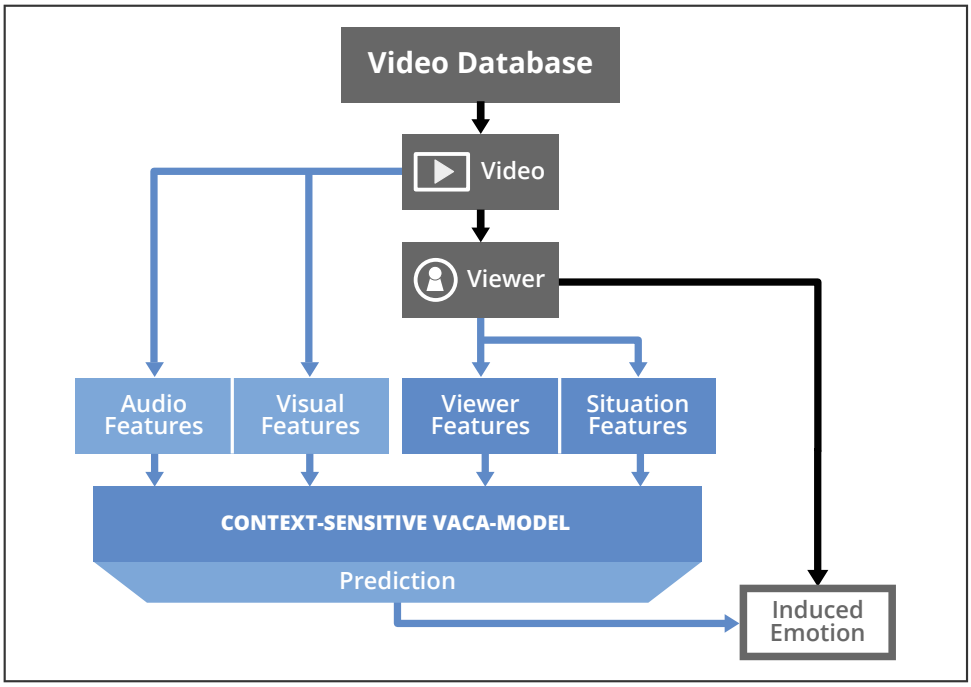


Figure 5.2: Context-Sensitive VACA approaches

theories underpinning prominent categorical representations (e.g., [34, 35]). Moreover, encoding emotions with a limited amount of categories may fail to capture differences at a level that is desired by applications, making dimensional schemes an attractive alternative for affective computing [36]. Perhaps the most prominent scheme for psychological investigations of media-induced emotions is the dimensional PAD framework [37]. It describes affective experiences in terms of the dimensions *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?). PAD has been used to code several prominent datasets for VACA research (e.g. *DEAP* [23] and *MAHNOB-HCI* [24]), as well as stimuli that are widely used for emotion elicitation in psychology (e.g. the IAPS image corpus [38]). As such, PAD may form a particularly good basis for incorporating insights from psychological research into VACA technology. While VACA practitioners sometimes discard the more comprehensive PAD scheme in favor of the simpler Pleasure-Arousal (PA) scheme [39], there are sound reasons for including dominance when modeling emotional responses, e.g., because it relates to central emotional appraisals [40].

Overall, there exists no consensus in VACA research about a particular way to represent emotional responses, resulting in substantial variation in approaches. While not a problem in itself, this makes it challenging to compare the psychological and technological implications of different computational research projects. Because of its relatively widespread use in VACA and the high degree of compatibility with psychological research, we adopt the PAD dimensional scheme for our data collection and modeling activities.

5.3. THE MEMENTOS DATASET

In this section, we describe the creation and relevant elements of a crowd-sourced dataset that we have collected for our experiments. The Mementos Dataset contains detailed information about viewers' responses to a series of music videos, including self-reports about their emotional responses and free-text descriptions of the memories that were triggered while watching them.

5.3.1. DATA COLLECTION PROCEDURE

We recruited 300 crowd-workers via *Amazon Mechanical Turk*, each receiving 6 *USD* for their participation. After providing their informed consent about participation and data usage, subjects filled in a survey with additional information about themselves and their situation (*Additional Context Measures*). Then each participant viewed a random selection of 7 stimuli from our pool of 42 music videos. After each video, we requested ratings for the emotions it had induced (*Induced Emotion*), followed by a questionnaire asking whether watching the video had triggered any personal memories. If this was the case, viewers were required to describe each of these memories with a short text (*Memory Descriptions*) and rate the feelings that they associate with these recollections (*Memory-Associated Affect*). This procedure resulted in a total of 2098 unique responses from the participants (49 to 50 for each video clip). Out of these, 978 responses from 260 unique participants involved the recollection of at least one personal memory. For the modeling activities reported in this chapter, we focus only on the subset of data that triggered personal memories. See *Table 5.1* for summary statistics.

5.3.2. VIDEO STIMULI

The dataset contains a selection of 42 music video segments from among a set of 150 that the researchers creating the DEAP corpus have previously evaluated for their induced affect [23]. We opted for these videos because of two reasons. The first one is the strong capacity of musical stimuli to trigger personal memories [11]. The second one is that the creators of the DEAP corpus collected PAD-ratings from multiple viewers for each evaluated video. We used these stimulus-wise ratings to balance the sample of videos that we selected for our study to produce both low and high variation responses across all three affective dimensions of the PAD-space. For example, for each video where viewers in the DEAP validation study displayed a high agreement in the pleasure dimension, we selected another one where agreement on pleasure was low).

5.3.3. SELF-REPORT MEASURES

Induced Emotions: We took self-reports for emotions that participants experienced while watching the video in the form of pleasure-, arousal- and dominance-ratings on a continuous scale in the interval of $[-1,+1]$. Participants provided them using the *AffectButton* instrument – a 2d-widget displaying an iconic facial expression that changes in response to users' mouse or touch interactions. Participants can select the facial expression that best fits their affect (see [41] for a detailed description and validation).

Table 5.1: Overview of Self-Reported Data for Viewers' whose Responses to Videos involved Personal Memories

Variable	Measurement	M (SD)	Min/Max
Induced Emotion	**Pleasure**	0.29 (0.53)	-1.00/1.00
$N = 978^*$	**Arousal**	-0.02 (0.80)	-1.00/1.00
	Dominance	0.25 (0.58)	-1.00/1.00
Mem.-asso. Affect	**Pleasure**	0.33 (0.53)	-1.00/1.00
$N = 978^*$	**Arousal**	0.05 (0.79)	-1.00/1.00
	Dominance	0.30 (0.58)	-1.00/1.00
Memory Descr.	**Word No.**	25.07 (15.45)	3/103
$N = 978^*$			
Personality	**Honesty**	2.68 (0.75)	0.50/4.00
$N = 260^+$	**Emotional.**	1.96 (0.78)	0.00/3.75
	Extravers.	2.54 (0.77)	0.00/4.00
	Agreeabl.	2.04 (0.67)	0.25/4.00
	Conscien.	2.62 (0.71)	0.50/4.00
	Openness	2.81 (0.66)	0.00/4.00
Mood	**Pleasure**	0.43 (0.40)	-0.66/1.00
$N = 260^+$	**Arousal**	-0.10 (0.78)	-1.00/1.00
	Dominance	0.38 (0.48)	-1.00/1.00
Demographics	**Age**	33.40 (6.06)	25/46
$N = 260^+$			
		N	N
	Gender	139 female	121 male
	National.	218 USA	42 Other

* Response-specific: measured once per response to a video
+ Viewers-specific: measured once per viewer

Memory-Associated Affect: We asked participants to rate the affective associations with each personal memory that a video has triggered using the AffectButton instrument. Participants could report and rate as many memories per video as they had experienced. However, only 51 out of 978 responses from viewers involved recollections of 2 or more memories. For these instances, we retained only the single memory for our modeling activities with the most intense associated affect in terms of PAD-ratings.

Memory Descriptions: Participants were requested to describe personal memories triggered in them with a short free-text description. We requested a response in English with a minimum length of three words. After filtering for multi-recollection responses (see the previous paragraph), we retained a 978 memory description.

Additional Context Measures: Participating viewers filled in a survey that provided us with additional person- and situation-specific information. This demographic information about viewers', i.e. their *age*, *gender* and *nationality*, as well as their personality traits in terms of the 6-factor *HEXACO*-model (Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to experience) collected with a brief questionnaire [42]. Finally, viewers' reported their mood at the time of participation (PAD ratings with the AffectButton).

5.4. INFLUENCE OF PERSONAL MEMORIES ON VIDEO-INDUCED EMOTIONS

In a previous study based on the Mementos dataset [14], we have demonstrated that 1. videos create more intensive and positive emotional responses when triggering personal memories in a viewer, and that 2. the occurrence of and affect associated with memories explains more variability in induced emotions than theoretically relevant viewer-specific measures, e.g., personality traits.

To further illustrate the relevance of personal memories as context for predicting individual viewers' experiences, we extend these earlier investigations here with a focused statistical analysis of only those responses that involve memories. Concretely, we quantify the variance explained by memory-associated affect in comparison to that of viewer-specific context variables captured in the dataset (see Demographics, Personality, and Mood in 5.1). For this purpose, we conduct a regression analysis with nested linear mixed-effects models targeting the emotions induced in viewers (one model per affective dimension: Pleasure (P), Arousal (A), and Dominance (D)). *Figure 5.3* shows the differences in explained variance by these models for 1. a baseline model (Vid), predicting only the video-specific average (i.e. the *Expected Emotion* in context-free VACA), 2. the combined contribution of *Demographics*, *Personality*, as well as *Mood* as additional predictors in the model ($+(De+Pe+Mo)$, and finally 3. the additional effect of *memory-associated affect* ($+Ma$).

We see that the combined viewer-specific measures account for a significant share of additional variance over the baseline (P: $\Delta R^2_m = .052$, $F(13,213.55) = 4.05$, $p < .001$; A: $\Delta R^2_m = .088$, $F(13,219.21) = 5.75$, $p < .001$; D: $\Delta R^2_m = .036$, $F(13,213.68) = 2.85$, $p < .001$). However, information about memory-associated affect explains a large share of unique

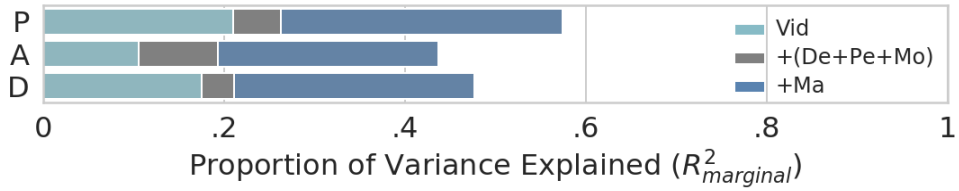


Figure 5.3: Comparison of the total variance explained by different context variables for dimensions of induced emotion. The video-specific average used in context-free VACA *(Vid)*. *Memory-associated affect (+Ma)* accounts for a greater share of additional uniquely explained variance than *demographics*, *personality* and *mood* combined *(De+Pe+Mo)*.

variation on top of that (P: $\Delta R^2_m = .364$, $F(16, 268.76) = 48.1$, $p < .001$; A: ($\Delta R^2_m = .332$, $F(16, 278.06) = 31.86$, $p < .001$); and D: ($\Delta R^2_m = .301$, $F(16, 263.22) = 263.22$, $p < .001$). In fact, the average amount of additional variance explained by memory-associated affect for affective dimensions is about *six times* higher. Together, these findings provide extremely compelling evidence for the exploitation of video-triggered memories as a form of situation-specific context for predictions in VACA.

5.5. PREDICTIVE MODELING

In this section, we outline the framework that we use as a proof of concept for integrating descriptions of personal memories triggered by a video as contextual information. We have made essential design choices, in how we collected our data to stay true to recognized affective theory and its relation to the task at hand. As a consequence, no state-of-the-art baseline exists from which we can compare our approach to others. We do *not* intend to provide significant extensions to the technological state-of-the-art. Instead, we provide a state-of-the-art baseline to investigate the contribution of memory descriptions such that we can better expose the nature of this novel task approach. Concretely, we describe the machine learning models and multimodal fusion strategies that we deploy for this task.

5.5.1. OVERVIEW

In line with previous work on VACA ([4]), we address modeling viewers' emotional responses as a regression problem. An important aspect of context-sensitive VACA technology is the integration of information from different sources into a single multimodal prediction model, i.e., video content and context features. Previous VACA work has repeatedly relied on early fusion to combine the information provided by different sources. Support Vector Machines are a popular type of machine learning algorithm for this purpose (e.g., [29, 31], especially when predictions take place in a regression setting (see [4]). We also consider the state of the art approaches for emotion prediction from text since we intend to exploit free-text descriptions of memories as context in predictions. Interestingly, state-of-the-art results for predicting the author's affective states from short texts (i.e., tweets) have been achieved by decision-level fusion. For instance, Duppada et al. used random forest regressors with a regularized linear model as a meta regressor different text feature-sets [43]. Motivated by this, we explore both early and late fusion approaches in our experiments. *Figure* 5.4 provides a graphical overview of the two machine

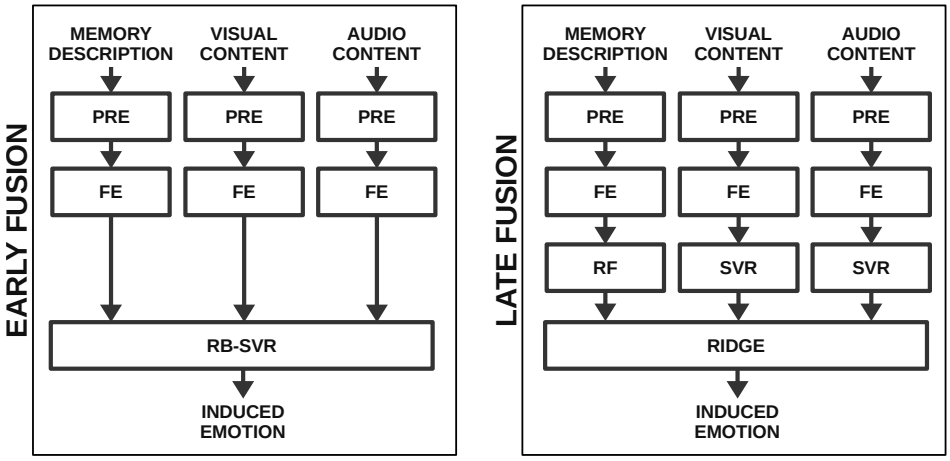


Figure 5.4: Overview of our approaches for predictive modeling and multimodal fusion. *PRE:* Preprocessing; *FE:* Feature Extraction; *SVR:* Support Vector Regression; *RF:* Random Forest Regression; *RIDGE:* L2-regularized linear model

learning pipelines that we deploy: an *Early Fusion-Approach*, consisting feature-level fusion of the different information sources, and a *Late Fusion-Approach*, using a meta-regressor to combine the predictions from modality-specific models via stacking. For all machine learning algorithms, we used the implementation provided by the *Scikit-Learn* python library [44] in Version 0.22.2.

Early Fusion-Approach: This pipeline consists of a series of pre-processing and feature extraction steps that are specific for memory, visual, and audio sources. The resulting feature vector is then concatenated, and fed to a Support Vector Regressor (SVR) with Radial Basis-Function (RBF) kernel for predictions.

Late Fusion-Approach Here, we also perform the same modality-specific pre-processing and feature extraction operations. We then rely on two separate SVRs with RBF kernel for predictions based on the audio or visual features, while, in line with [43], we use a Random Forest Regressor for predictions based on free text memory descriptions. We combine the output of the modality-specific models with an L2-regularized linear regression model (Ridge regression).

5.5.2. STIMULUS VIDEO PROCESSING

To represent the audio content of the music videos, we used the software *openSMILE* in the configuration *"emobase2010"* for feature extraction. It derives low-level descriptors from raw audio signals in a windowed fashion and aggregates them statistically, resulting in a 1582-dimensional feature vector (see [45] for a detailed description). This feature set is widely used for audio representation in VACA research as a baseline approach in benchmarking challenges [46].

For visual representation of the stimulus video, one frame is extracted every second. For each frame, we then extract three types of visual features:

Theory-inspired Descriptors: Work on affective visual content analysis has developed features that were specifically engineered to capture the affective properties of images. These descriptors often are inspired by findings from psychological research or art-theoretic concepts. We use the set of descriptors developed by Machajdik & Hanburry [47], as well as those of Bhattacharya et al. [48] to characterize each of the extracted video frames (resulting in a 271-dimensional feature vector). This combination has been used in context-sensitive VACA work before [31].

Deep Visual Descriptors: Deep neural networks form an essential part of modern approaches to visual content analysis and computer vision. Instead of relying on engineered descriptors of visual input, these models learn effective and reusable representations for prediction tasks directly from visual training data. We use the activation of the FC1-layer of a pre-trained VGG16 network [49] from the Keras framework for python [50] for this purpose (resulting in a feature vector with 4096 dimensions). This representation has been used for visual representation in VACA research as a baseline in benchmarking challenges [46].

Visual Sentiment Descriptors: Classifiers that identify the presence of *Adjective-Noun Pairs (ANPs)* in images have been successfully used as high-level descriptors in VACA approaches (e.g., [29, 31]). ANPs consist of labels denoting objects or persons identified in an image, coupled with an affective attribute (e.g., "beautiful house"). We use the class-probabilities assigned by the *DeepSentiBank* Network [51] for any of the ANPs in its ontology as descriptors for the content of video frames (4342-dimensional feature vector).

We concatenate these different feature sets into a combined 8709-dimensional vector to represent the visual content of individual extracted video frames. We compute these for each of the frames extracted from a video, and then take the dimension-wise average across them to produce a single 8709-dimensional representation of the entire stimulus video's visual content.

5.5.3. MEMORY DESCRIPTION PROCESSING

We preprocess memory descriptions by replacing references to specific years or decades (e.g. "1990", or "the 90s") with generic terms (e.g. "that year" or "that decade"). Additionally, we replace any numbers with 0 and expand all contractions present in participants' descriptions (e.g. "can't" is transformed into "cannot"). To model the affective impact of personal memories we rely on features that have proven successful in state-of-the-art work modeling emotional states from social media text in a regression setting [20]: 1. *Lexical Features* and 2. *Word Embeddings*.

Lexical Features: We generate these features by parsing memory descriptions into word-level tokens, for which we then retrieve associated affective ratings from a wide variety of affective dictionaries. To account for differences between words used in memory descriptions and the form in which they are typically indexed in lexica, we apply lemmatization

before the lookup to remove inflections. The combination of the dictionaries that we initially selected for feature extraction [52–62] was demonstrated to contribute to state-of-the-art performance for affective text regression tasks [43]. We extended this list by a recent addition that provides word-level ratings for Pleasure, Arousal, and Dominance [20]. We aggregate the associated word-level ratings from each dictionary by averaging the extracted features for each word token in a memory description. Additionally, we include the sentiment scores provided by the *VADER* model [63] when applied to an entire memory description. It combines an empirically collected sentiment lexicon with a set of rule-based processing steps to score the affective valence of text. Together this results in a 130-dimensional vector of lexical features for each description.

Word Embeddings: We leveraged two pre-trained word embedding-models to represent each word in the memory description texts as a real-valued feature vector: *(1) Word2Vec*-model pre-trained on the *Google News dataset*, resulting in a 300-dimensional feature vector when applied to a word, and *(2)* a *GloVE*-model [64] pre-trained on the *Wikipedia 2014 and the Gigaword 5 corpora*. It encodes individual words as a 200-dimensional feature vector. For both implementations we rely on the *Gensim*-library for python [65]. To generate a representation for the entire memory description from these word-level embeddings, vectors of both types are concatenated and averaged for the entire description, resulting in a single 500-dimensional feature vector for each memory.

5.6. EMPIRICAL INVESTIGATIONS

We conduct two experiments in an ablation setting of our early and late fusion approaches when predicting the video-induced emotions of viewers in the Mementos Dataset. In *Experiment 1*, we assess the feasibility of predicting viewers' emotional responses to a video from text describing the memories that it has triggered in them. *Experiment 2* quantifies the relative contribution of memory descriptions when used for predicting emotional responses alongside the audiovisual content of videos.

5.6.1. EXPERIMENTAL SETUP AND EVALUATION

We use a nested 5-Fold-Leave-Persons-Out Cross-Validation for training and evaluation of our early and late fusion approaches. In this procedure, folds are created such that no data from the same individual is spread across training and validation. In the outer loop of the nested cross-validation, we split the entire dataset into 5 folds, from which we hold out a single fold for testing the final performance of selected models. In the inner loop, the remaining 4 folds are used for selecting hyperparameters for the machine learning models with a grid search.

5.6.2. RESULTS AND ANALYSIS

EXPERIMENT 1 – USING MEMORY DESCRIPTIONS FOR PREDICTION

We show the average performance of our early and late fusion approaches when only provided with viewers' memory descriptions as input in *Figure 5.5*. These findings indicate that text descriptions explain a significant portion of the variance in induced pleasure and, to a lesser degree, dominance (i.e., their $AvgR^2 > 0$). However, the performance of

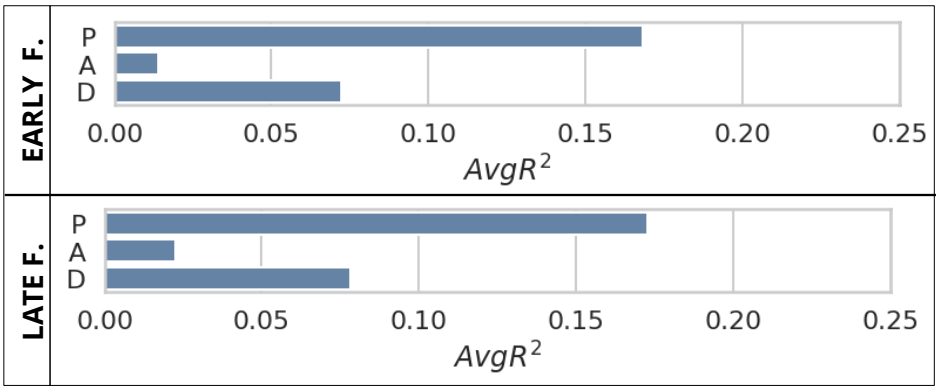


Figure 5.5: Results for Experiment 1 — Average Test-Performance for Early and Late Fusion Approaches for predicting viewers' induced *Pleasure (P)*, *Arousal (A)*, and *Dominance (D)* when using only Memory Descriptions.

our models is much lower for arousal.

To explore whether this decrease is a result of our modeling choices, we investigate how well humans can infer affective information from the memory descriptions in our dataset. For this purpose, two raters manually annotated a random selection of 150 descriptions with two kinds of affective evaluations for pleasure, arousal, and dominance: *(1) Perceived Conveyed Affect (PCA)* of the text, and *(2)* the *Inferred Affective Experience (IAX)* of the author.

For PCA ratings, the annotators answer the question *"What feelings does this text express?"*. We instruct them only to consider emotions or feelings that are explicitly described by the authors, e.g., by using emotion words like "love" or "hate". Performance on this task will provide us with insights about 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 the different task formulation is to encourage annotators to draw on their cultural background and experience to infer implicit emotional meaning from the descriptions. Such inferences are a vital component of human emotion perception [66], and performance on this task shows the degree to which the texts facilitate them. Raters provide PAD annotations with the widely adopted and validated *Self-Assessment Manikin* instrument [67]. Based on this information, we assess the *correspondence* of the two annotators' ratings with viewers' self-reported ratings for memory-associated affect (MA), and the *reliability* with which they were able to do so. For this, we calculate Pearson correlations between the average PCA/IAX ratings of both annotators and viewers' MA ratings to measure correspondence. Similarly, we calculated the correlations between the PCA/IAX ratings between both raters as a measure of reliability and agreement. Results for both are listed in *Table 5.2*, where statistical significance was determined using a clustered bootstrap procedure (10000 repetitions) to account for nesting of descriptions in participants.

Our findings show that annotators' ratings for pleasure in the PCA and IAX tasks both

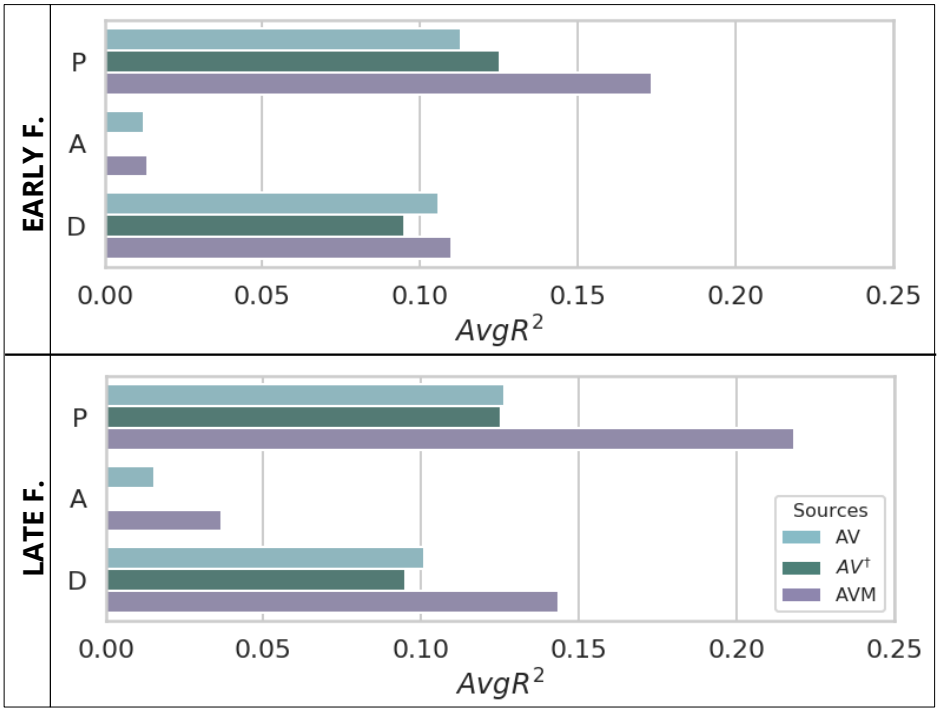


Figure 5.6: Results for Experiment 2 — Average Test-Performance for Early and Late Fusion Approaches when predicting induced *Pleasure (P)*, *Arousal (A)*, and *Dominance (D) using different sources. AV : audio-visual tracks; AV† : Model approximating optimal results for Context-Free VACA using the video-specific mean of the ground truth; AVM : audio-visual tracks combined with memories.*

highly agree with viewers' experienced emotions, as well as each other. This pattern is still present – albeit less strongly pronounced – for dominance. However, for arousal, annotators' judgments correspond much less with viewers' MA ratings. Moreover, annotators also tend to disagree much more with each other. On average, both correspondence and reliability of judgments are higher in the IAX task than for the PCA task. This result confirms our hypothesis that simulating the authors' state of mind helps human annotators. However, levels of performance do still not reach those displayed for pleasure or dominance. This finding points towards an inherently greater difficulty for recognizing arousal from text, rather than to a particular weakness in our modeling approach. A likely explanation is that descriptions contain few explicit or implicit expressions of an author's arousal, making it challenging for both humans and automatic text analysis to perform well.

EXPERIMENT 2: COMBINING DIRECT VACA AND MEMORY CONTEXT

Figure 5.6 shows the performance displayed by our approaches when having access to either 1. only videos' audiovisual data (*AV*), or 2. a combination of both audiovisual data and memory descriptions (*AVM*). In addition, we also list the performance of a model

Table 5.2: Human Annotators' Performance for Affective Interpretation of Memory Descriptions (Pearson Correlations)

	Correspondence			Reliability		
	P	A	D	P	A	D
PCA	.586***	.147	.389***	.773***	.181	.54***
IAX	.555***	.26**	.433***	.859***	.314***	.578***

* $p <= .05$; ** $p <= .01$; *** $p <= .001$

that predicts the video-wise mean for affective dimensions of induced emotions from a sample ($AV\dagger$). This measure indicates the best-case performance that a context-free VACA model can provide for the current dataset if it always makes the correct prediction of the "Expected Emotion" for a video.

Our analysis shows that memory descriptions can provide substantial additional information about viewers' affective responses, independent of fusion strategy. Overall, improvements when giving models access to memory descriptions are most pronounced for predictions of pleasure (Early Fusion: $\Delta AvgR^2 = +.059$; Late Fusion: $\Delta AvgR^2 = +.092$;). This evidence is consistent with the findings of experiment 1, which demonstrates the greater ability of our models for predicting pleasure. Similarly, the performance for arousal remains poor across the board, even when using audiovisual features only. Despite the comparably simple modeling approach that we have deployed, the performance of our models using only audiovisual features (AV) approximates that offered by an ideal context-free VACA model ($AV\dagger$) (absolute differences in performance averaged across affective dimensions – Early Fusion: $\Delta AvgR^2 = .003$; Late Fusion: $\Delta AvgR^2 = .006$). Finally, we observe that the late fusion approach displays a two times greater increase in performance when provided with memory descriptions than the early fusion approach (difference in performance gains averaged across affective dimensions – Early: $\Delta AvgR^2 = +.021$; Late: $\Delta AvgR^2 = +.054$). This result highlights that despite the efficacy of early fusion in classic VACA, research should not rule out late fusion for context-sensitive approaches.

5.7. DISCUSSION

Our empirical investigations demonstrate that it is both feasible to use viewers' self-reported memory descriptions for predicting emotional experience and that doing so provides a valuable source of context for personalized predictions in video affective content analysis. Particularly for pleasure, the automatic analysis of memory descriptions explains variation that is substantially above and beyond that of a video's audio and visual content. Surprisingly, none of our models provided substantial insights into viewers' arousal. Our findings that even humans struggle to reliably and accurately infer arousal from memories offer a possible explanation. Nevertheless, the weak performance for arousal when using stimulus features is surprising, since previous research achieved performance of arousal predictions comparable to pleasure or dominance (e.g., [23, 68, 69]). A reason for this might be that we did not filter our stimuli based on the variability of responses they elicited before modeling. Especially for arousal, our viewers reported widely different levels of arousal to the same video. This fluctuation might make it harder for algorithms to learn directly from a video's audiovisual content. Similarly, the stimulus-

specific mean as "Expected Emotion" might not be a good approximation for these cases. This last finding underlines the limited capacity of purely context-free predictions to reflect viewers' individual experiences of video stimuli accurately.

Overall, the studies demonstrate that descriptions of video-triggered personal memories are a viable and useful resource for making personalized predictions. However, there are some limitations to our findings. First, the memory descriptions that we analyzed may differ from those that viewers would create in-the-wild, such as on social media platforms, or during a conversation. Second, we have collected them from paid crowd-workers, who provided them with the full knowledge that their identity remains protected. Under these circumstances, participants may have been willing to provide more detailed and candid accounts of their memories than they otherwise would have. Future work could expand on our findings by compiling corpora that capture how people describe media-evoked recollections under more natural conditions in the wild. An effective way to achieve this might be to elicit recollections over more extended periods with specifically developed applications, e.g., via social media [70], or through conversations with interactive intelligent agents [71]. Moreover, research could explore technological efforts to identify and extract memory descriptions from generally available data – e.g., social media posts. – automatically.

However, relying on the availability of explicit descriptions can only be a first step for improving predictions in real-world settings. For many use-cases of VACA, such descriptions may not be available at prediction time, as viewers can only describe the memories triggered in them by a video after they have already been exposed to it, but not before. To anticipate the effect of memories on viewers beforehand – e.g., to personalize recommendations of unseen content –, models will have to estimate when memories are triggered and what their content will be. While these are undoubtedly challenging tasks, progress towards achieving them in automated systems seems feasible. A first step could be to explore how well the use of already obtained text-based memory descriptions generalizes to new, but related stimuli, e.g. music videos from the same artist. Additionally, Dudzik et al. [72] argues that existing research from ubiquitous computing and cognitive modeling offers numerous starting points for modeling memory processes in adaptive media technology. For example, they propose that modeling a user's attentional focus from sensor data might be a way to identify situations and stimuli that are likely to trigger memories. Given the substantial role that personal memories play for induced emotions, such research topics should be actively pursued in context-sensitive VACA. Naturally, this requires access to rich corpora of data that facilitate such investigations.

5.8. SUMMARY AND CONCLUSION

Video-affective Content Analysis (VACA) has traditionally operated under the assumption that videos possess a more or less objective emotional meaning, existing across different viewers and independent of the situation under which they are experienced. However, the emotional impact of videos in the real world is highly subjective and varies greatly with the situation. Consequently, not accounting for context limits the ability of predictions to reflect emotional responses accurately. Contemporary video content is consumed by an increasingly diverse, global-spanning community and in a broad variety of circumstances. Given these developments, research on context-sensitive approaches to VACA seems vital

for predictions to be of use to media applications.

Personal memories that are triggered in viewers form a highly situation-specific form of context, shaping individual emotional responses to the video. The two empirical investigations that we describe in this chapter show the feasibility of using free-text descriptions to account for this influence in automated predictions. Moreover, our findings demonstrate that combining this approach with an analysis of a video's audiovisual content can provide significant performance benefits. As such, when memory descriptions are available, they offer a powerful form of context for personalizing predictions in VACA. Because people tend to talk about their memories with each other, automatic approaches can feasibly extract such descriptions from communications on social media or face-to-face interactions. Nevertheless, the investigations described in this chapter only form a first step towards accounting for the influence of memories in automatic predictions. Future investigations should more generally explore modeling the occurrence (when?), content (what?), and influence (what does it do?) of personal memories in VACA predictions to support media applications in the real world.

REFERENCES

[1] A. Bartsch, *Emotional Gratification in Entertainment Experience. Why Viewers of Movies and Television Series Find it Rewarding to Experience Emotions,* Media Psychology **15**, 267 (2012).

[2] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, *Affective Video Content Analysis: A Multidisciplinary Insight,* IEEE Transactions on Affective Computing **9**, 396 (2018).

[3] A. Hanjalic and Li-Qun Xu, *Affective video content representation and modeling,* IEEE Transactions on Multimedia **7**, 143 (2005).

[4] S. Wang and Q. Ji, *Video Affective Content Analysis: A Survey of State-of-the-Art Methods,* IEEE Transactions on Affective Computing **6**, 410 (2015).

[5] K. H. Greenaway, E. K. Kalokerinos, and L. A. Williams, *Context is Everything (in Emotion Research),* Social and Personality Psychology Compass **12**, e12393 (2018).

[6] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, *Corpus Development for Affective Video Indexing,* IEEE Transactions on Multimedia **16**, 1075 (2014), arXiv:1211.5492 .

[7] Z. Hammal and M. T. Suarez, *Towards context based affective computing introduction to the third international CBAR 2015 workshop,* in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (IEEE, 2015) pp. 1–2.

[8] B. Dudzik, M.-P. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong, *Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases,* in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 206–212.

[9] D. G. McDonald, M. A. Sarge, S.-F. Lin, J. G. Collier, and B. Potocki, *A Role for the Self: Media Content as Triggers for Involuntary Autobiographical Memories,* Communication Research **42**, 3 (2015).

[10] A. M. Belfi, B. Karlan, and D. Tranel, *Music evokes vivid autobiographical memories,* Memory **24**, 979 (2016).

[11] P. Janata, S. T. Tomic, and S. K. Rakowski, *Characterisation of music-evoked autobiographical memories,* Memory **15**, 845 (2007).

[12] C. Mills and S. D'Mello, *On the Validity of the Autobiographical Emotional Memory Task for Emotion Induction,* PLoS ONE **9**, e95837 (2014).

[13] H. Baumgartner, *Remembrance of things past: Music, autobiographical memory, and emotion,* Advances in Consumer Research **19**, 613 (1992).

[14] Authors, *Anonymous for Peer-Review,* in *Forthcoming. See document in supplement files for anonymized version.* (2020).

[15] D. Cosley, V. S. Sosik, J. Schultz, S. T. Peesapati, S. Lee, T. Peesapati, and S. Lee, *Experiences With Designing Tools for Everyday Reminiscing,* Human–Computer Interaction **Volume 27**, 175 (2012).

[16] B. Rimé, S. Corsini, and G. Herbette, *Emotion, verbal expression, and the social sharing of emotion,* The verbal communication of emotions: Interdisciplinary perspectives , 185 (2002).

[17] W. R. Walker, J. J. Skowronski, J. A. Gibbons, R. J. Vogl, and T. D. Ritchie, *Why people rehearse their memories: Frequency of use and relations to the intensity of emotions associated with autobiographical memories,* Memory **17**, 760 (2009).

[18] S. Bluck, N. Alea, T. Habermas, and D. C. Rubin, *A TALE of Three Functions: The Self–Reported Uses of Autobiographical Memory,* Social Cognition **23**, 91 (2005).

[19] B. Caci, M. Cardaci, and S. Miceli, *Autobiographical memory, personality, and Facebook mementos,* Europe's Journal of Psychology **15**, 614 (2019).

[20] S. Mohammad, *Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words,* in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018) pp. 174–184.

[21] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, *SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text,* in *Proceedings of the 13th International Workshop on Semantic Evaluation* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019) pp. 39–48.

[22] 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 *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 697–703.

[23] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *DEAP: A Database for Emotion Analysis ;Using Physiological Signals,* IEEE Transactions on Affective Computing **3**, 18 (2012).

[24] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *A multimodal database for affect recognition and implicit tagging,* IEEE Transactions on Affective Computing **3**, 42 (2012).

[25] S. K. D'mello and J. Kory, *A Review and Meta-Analysis of Multimodal Affect Detection Systems,* ACM Computing Surveys **47**, 1 (2015).

[26] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian, *Evaluating content-centric vs. user-centric ad affect recognition,* in *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017* (ACM Press, New York, New York, USA, 2017) pp. 402–410, arXiv:1709.01684 .

[27] M. Soleymani and M. Pantic, *Human-centered implicit tagging: Overview and perspectives,* in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* (IEEE, 2012) pp. 3304–3309.

[28] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, *DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses,* IEEE Transactions on Affective Computing **6**, 209 (2015).

[29] D. McDuff and M. Soleymani, *Large-scale Affective Content Analysis: Combining Media Content Features and Facial Reactions,* in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (IEEE, 2017) pp. 339–345.

[30] 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 *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2015) pp. 236–242.

[31] M. J. Scott, S. C. Guntuku, W. Lin, and G. Ghinea, *Do Personality and Culture Influence Perceived Video Quality and Enjoyment?* IEEE Transactions on Multimedia **18**, 1796 (2016).

[32] Y. Zhu, I. Heynderickx, and J. A. Redi, *Understanding the role of social context and user factors in video Quality of Experience,* Computers in Human Behavior **49**, 412 (2015).

[33] S. Wang, Y. Zhu, G. Wu, and Q. Ji, *Hybrid video emotional tagging using users' EEG and video content,* Multimedia Tools and Applications **72**, 1257 (2014).

[34] A. Ortony and T. J. Turner, *What's basic about basic emotions?* Psychological Review **97**, 315 (1990).

[35] L. F. Barrett, *Solving the Emotion Paradox: Categorization and the Experience of Emotion,* Personality and Social Psychology Review **10**, 20 (2006).

[36] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, *Emotion representation, analysis and synthesis in continuous space: A survey,* in *Face and Gesture 2011* (IEEE, 2011) pp. 827–834.

[37] A. Mehrabian, *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament,* Current Psychology **14**, 261 (1996).

[38] P. J. Lang, M. M. Bradley, B. N. Cuthbert, and Others, *International affective picture system (IAPS): Technical manual and affective ratings,* NIMH Center for the Study of Emotion and Attention **1**, 39 (1997).

[39] J. A. Russell, *A circumplex model of affect.* Journal of Personality and Social Psychology **39**, 1161 (1980).

[40] J. Broekens, *In Defense of Dominance,* International Journal of Synthetic Emotions **3**, 33 (2012).

[41] J. Broekens and W.-P. Brinkman, *AffectButton: A method for reliable and valid affective self-report,* International Journal of Human-Computer Studies **71**, 641 (2013).

[42] R. E. de Vries, *The 24-item Brief HEXACO Inventory (BHI),* Journal of Research in Personality **47**, 871 (2013).

[43] V. Duppada, R. Jain, and S. Hiray, *SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets,* Proceedings of The 12th International Workshop on Semantic Evaluation , 18 (2018), arXiv:1804.06137 .

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and Others, *Scikit-learn: Machine learning in Python,* Journal of machine learning research **12**, 2825 (2011).

[45] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, *The INTERSPEECH 2010 paralinguistic challenge,* in *Eleventh Annual Conference of the International Speech Communication Association* (2010).

[46] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, Z. Xiao, and M. Sjöberg, *The MediaEval 2018 emotional impact of Movies task,* CEUR Workshop Proceedings , 1.

[47] J. Machajdik and A. Hanbury, *Affective image classification using features inspired by psychology and art theory,* in *Proceedings of the international conference on Multimedia - MM '10* (ACM Press, New York, New York, USA, 2010) p. 83.

[48] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, *Towards a comprehensive computational model foraesthetic assessment of videos,* in *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, 3 (ACM Press, New York, New York, USA, 2013) pp. 361–364.

[49] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition,* in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2015) arXiv:1409.1556 .

[50] F. Chollet and Others, *Keras,* \url{https://keras.io} (2015).

[51] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, *DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks,* (2014), arXiv:1410.8586 .

[52] M. Hu and B. Liu, *Mining and summarizing customer reviews,* in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04* (ACM Press, New York, New York, USA, 2004) p. 168.

[53] S. Baccianella, A. Esuli, and F. Sebastiani, *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.* in *Lrec*, Vol. 10 (2010) pp. 2200–2204.

[54] S. Mohammad and P. Turney, *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon,* in *Proceedings of the {NAACL} {HLT} 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (Association for Computational Linguistics, Los Angeles, CA, 2010) pp. 26–34.

[55] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, *Sentiment strength detection in short informal text,* Journal of the American Society for Information Science and Technology **61**, 2544 (2010).

[56] Y. R. Tausczik and J. W. Pennebaker, *The psychological meaning of words: LIWC and computerized text analysis methods,* (2010).

[57] F. Å. Nielsen, *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,* CEUR Workshop Proceedings **718**, 93 (2011), arXiv:1103.2903 .

[58] S. M. Mohammad, S. Kiritchenko, and X. Zhu, *NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets,* *SEM 2013 - 2nd Joint Conference on Lexical and Computational Semantics **2**, 321 (2013), arXiv:1308.6242 .

[59] Y. Choi and J. Wiebe, *+/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference,* in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014) pp. 1181–1191.

[60] S. M. Mohammad and S. Kiritchenko, *Using Hashtags to Capture Fine Emotion Categories from Tweets,* Computational Intelligence **31**, 301 (2015).

[61] F. Bravo-Marquez, E. Frank, S. M. Mohammad, and B. Pfahringer, *Determining Word-Emotion Associations from Tweets by Multi-label Classification,* in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (IEEE, 2016) pp. 536–539.

[62] S. M. Mohammad, *Word Affect Intensities,* (2017).

[63] C. J. Hutto and E. Gilbert, *VADER: A parsimonious rule-based model for sentiment analysis of social media text,* in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (2014) pp. 216–225.

[64] J. Pennington, R. Socher, and C. Manning, *Glove: Global Vectors for Word Representation,* in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014) pp. 1532–1543.

[65] R. Řehůřek and P. Sojka, *Software Framework for Topic Modelling with Large Corpora,* in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA, Valletta, Malta, 2010) pp. 45–50.

[66] L. F. Barrett, B. Mesquita, and M. Gendron, *Context in Emotion Perception,* Current Directions in Psychological Science **20**, 286 (2011).

[67] M. M. Bradley and P. J. Lang, *Measuring emotion: The self-assessment manikin and the semantic differential,* Journal of Behavior Therapy and Experimental Psychiatry **25**, 49 (1994).

[68] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, *Multimedia content analysis for emotional characterization of music video clips,* Eurasip Journal on Image and Video Processing **2013**, 26 (2013).

[69] Y. Baveye, E. Dellandrea, C. Chamaret, and Liming Chen, *LIRIS-ACCEDE: A Video Database for Affective Content Analysis,* IEEE Transactions on Affective Computing **6**, 43 (2015).

[70] D. Cosley, V. Schwanda, S. T. Peesapaty, J. Schultz, and J. Baxter, *Experiences with a publicly deployed tool for reminiscing,* in *Proc. First Int'l Workshop on Reminiscence Systems* (2009) pp. 31–36.

[71] M. M. Peeters, M. Harbers, and M. A. Neerincx, *Designing a personal music assistant that enhances the social, cognitive, and affective experiences of people with dementia,* Computers in Human Behavior **63**, 727 (2016).

[72] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, *Artificial Empathic Memory,* in *Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions - EE-USAD'18* (ACM Press, New York, New York, USA, 2018) pp. 1–8.

6

Personal Memory Appraisal as Context for Facial Behavior Analysis

This chapter is based on: Dudzik, B., Broekens, J., Neerincx, M., & Hung, H. (2020). *Exploring Personal Memories and Video Content as Context for Facial Behavior in Predictions of Video-Induced Emotions.* Proceedings of the 2020 International Conference on Multimodal Interaction, 10(20), 153–162.

ABSTRACT

Empirical evidence suggests that the emotional meaning of facial behavior in isolation is often ambiguous in real-world conditions. While humans complement interpretations of others' faces with additional reasoning about context, automated approaches rarely display such context-sensitivity. Empirical findings indicate that the personal memories triggered by videos are crucial for predicting viewers' emotional response to such videos — in some cases, even more so than the video's audiovisual content. In this chapter, we explore the benefits of personal memories as context for facial behavior analysis. We conduct a series of multimodal machine learning experiments combining the automatic analysis of video-viewers' faces with that of two types of context information for affective predictions: (1) self-reported free-text descriptions of triggered memories and (2) a video's audiovisual content. Our results demonstrate that both sources of context provide models with information about variation in viewers' affective responses that complement facial analysis and each other.

6.1. INTRODUCTION

The capacity of video content to induce specific emotions – e.g., feelings of joy, sadness, and even disgust – is an essential motivator for people to engage with them [1]. For this reason, research is exploring the development of intelligent media technologies that can recognize and learn from users' emotional responses, e.g., to facilitate personalized content recommendations [2].

The automatic analysis of facial behavior is traditionally an essential method for automatic affect detection [3], including the recognition of emotional responses to video stimuli (e.g., [4–6]). However, findings from empirical psychology increasingly reveal that the face offers only limited insight into a person's feelings outside of artificially created laboratory settings [7]. Rather than displaying a clear correspondence with a persons' affective state, numerous studies have demonstrated that the emotional meaning of spontaneous facial behavior in the real world is often ambiguous and highly variable [8]. These findings have direct consequences for the performance of automatic systems that analyze faces for detecting affective states of users. Studies evaluating commercially available software have also revealed challenges for predictions to correspond with self-reported affect [9], as well as the perceptions of third-party observers [10].

Instead of relying solely on interpreting behavioral cues, human perceivers draw on contextual knowledge about the background and present situation of an observed person to reason about potential influences on their feelings [11–13]. The insights gained by this act of emotional perspective-taking can complement any information offered by behavior in isolation, thereby enabling an observer to make accurate inferences even for ambiguous cases (e.g., [14]). However, context-sensitive approaches remain under-explored in automatic affect detection [15], despite researchers generally acknowledging their potential [16–18]. Likely causes for this neglect are the substantial challenges involved in 1. identifying relevant contextual influences for emotional responses in an application setting, as well as 2. developing technical solutions that provide automatic systems with an awareness of them [19]. Overcoming these challenges requires systematic exploration of person- and situation-specific influences in computational modeling activities [15]

informed by findings from the social sciences [8]. Compared to emotional responses in general, situations in which video stimuli are consumed by an individual provide a more constrained scenario for the exploration of relevant contextual influences. For example, it is reasonable to assume that the video's content has a strong influence on viewers' emotional responses and that its analysis can aid automatic affect detection (e.g., [6]). However, numerous other important influences exist [16].

In this chapter, we contribute to the development of context-sensitive recognition of video-induced emotions by demonstrating the benefits of accounting for video-triggered personal memories as additional context in automated predictions. Empirical findings indicate that media are both powerful cues for personal memories in observers [20, 21] and that the evoked memories are a powerful causal influence on emotional responses [22]. Moreover, the feelings associated with any memories triggered by a video in this way closely relate to its overall emotional impact [23], i.e., positive memories lead to a more positive response to a video. These findings indicate that information about the content of personal memories associated with a particular video can provide insights into its emotional impact. Moreover, because triggered memories constitute a contextual influence shaping or even causing emotions during video-viewing, they may also facilitate inferences when viewers do not overtly express their feelings. For this reason, accounting for the occurrence and emotional significance of personal memories in automated predictions has a strong potential to complement the analysis of viewers' behaviors.

One possible way to achieve this is through the automatic analysis of text or speech data in which individuals explicitly describe memories triggered in them while watching a video. Findings indicate that people frequently disclose memories from their personal lives to others [24, 25], for example, in service of social bonding or emotion regulation processes [26]. There is evidence that people share memories for similar reasons on social media [27], and that they readily describe memories triggered in them by social media content [28]. Additionally, research is extensively exploring both the automatic affective analysis of text-data from social media [29], and that of face-to-face dialog [30]. Building on existing work in this area, we have previously established that self-reported free-text descriptions video-triggered memories can be successfully used for predictions, improving performance over automatic analysis of video content in isolation [31]. Motivated by this, we present here the following contributions to the field:

- We conduct a series of multimodal machine learning experiments using a dataset capturing peoples' emotional responses to music videos to predict induced emotions based on analysis of viewers' facial behavior, in combination with memory content and video content. Our findings demonstrate that incorporating information about both forms of context improves predictive performance.
- Using statistical analysis, we establish that video content and memory descriptions provide strong complementary information about viewers' experience of pleasure and dominance, but not arousal. Memories emerged as the best overall source of information for predictions.
- We outline opportunities for future research to account more comprehensively for memory-influences in automated affect detection and potential benefits for applications.

In the remaining chapter, we first discuss related work on context-sensitive automatic affect detection and motivate our choice of affect representation. Then we describe the dataset and approach for predictive modeling used in our empirical investigations. We conclude with a detailed analysis and discussion of our findings.

6.2. BACKGROUND AND RELATED WORK

6.2.1. CONTEXT IN AFFECT DETECTION

In the following, we provide a brief discussion of some types of contextual information that psychological research has identified as relevant for human emotion perception, and how existing technological research addresses it. When interpreting another person's facial expression, humans rely on sensory information present in the scene surrounding it and previous knowledge and experiences that they bring into the scene [11]. A basic form of sensory information is other behavioral signals and cues, e.g., body posture and gestures [12]. Such *cross-behavior context* has been extensively explored in multimodal analysis approaches, especially with speech as an added modality, typically showing performance improvements [3]. Additionally, human perceivers rarely observe (facial) behavior in the form of isolated snapshots but instead as firmly embedded in a *temporal context*. Exploiting such temporal dependencies of behavioral data is conceptually relatively straight forward. It is the topic of a substantial amount of technological research in automated affect detection (see Rouast et al. [32] for an overview of recent deep learning-based approaches).

The observable scene surrounding another person can be an essential source of information for inferences of their emotional state [12]. Importantly, it forms the foundation for perceivers to reason about aspects of the *situation-specific context* that causes or shapes the other's response. Such information about triggering events has a strong role in interpreting facial behavior [13]. Affective detection work has only tentatively explored this aspect because it is conceptually challenging to translate into automatic systems and generally lacks available corpora for modeling [15]. Notably, however, Kosti et al. [33] demonstrate the benefits offered by the visual scene as context in a large-scale approach for image-based affect detection. In contrast to generic affect detection, video-induced emotion recognition provides a more constrained scenario regarding situation-specific contextual influences. For example, due to the nature of the task, it is reasonable to assume that the eliciting video stimulus's content is an essential driver of emotional responses. For this reason, several multimodal approaches have combined analysis of it with that of facial behavior (e.g., [5, 6, 34]). Similarly, when viewing occurs in a social setting with multiple persons, looking at other viewers' behavior might provide context for predictions in computational models [35].

To summarize: while individual research projects model relevant influences on video-induced emotions, accounting for context is not yet pursued systematically. Notably, cognitive influences during consumption, such as elicited personal memories, have not yet been explored in computational work.

6.2.2. REPRESENTING AFFECTIVE STATES FOR DETECTION

A challenging aspect of developing systems for automatic affect detection is the conceptualization of the targeted states [3], including a formal scheme according to which the system characterizes and distinguishes between affective states – i.e., an *affect representation*. Affective Computing research has traditionally relied on two types of schemes to represent emotions for recognition: categorical and dimensional frameworks. Categorical schemes classify emotions in terms of a set of discrete states, such as happiness or anger. On the other hand, dimensional schemes describe human affect in terms of points in a continuous, multidimensional space, where each dimension is supposed to capture an aspect that is crucial for discriminating between different feelings. Traditionally, face-based affect detection has favored categorical schemes, since the underlying psychological theories postulate a strong connection between certain prototypical facial expressions and feelings. However, empirical evidence suggests that these associations are highly context-dependent and overall comparatively weak outside of laboratory studies [8]. Moreover, categorical schemes have been considered as not expressive enough to capture the degree of nuance relevant for some real-world applications, leading researchers to increasingly favor dimensional schemes [36]. A widely used dimensional framework is *Pleasure-Arousal-Dominance (PAD)* [37]. It describes emotions in terms of the three dimensions *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?). Because of its popularity for both psychological research (e.g. the widely used IAPS corpus [38]) and automatic affect detection (e.g. *DEAP* [34], or *EMOTIC* [39] corpora), we use it to represent emotions in our modeling activities. Additionally, PAD captures dominance (in contrast to only Pleasure and Arousal), which can be linked to emotional appraisals important for applications [40].

6.3. DATASET

In this section, we provide an overview of a corpus collected via crowd-sourcing for our modeling activities. It captures people's responses to music videos that they are watching on their electronic devices, including audiovisual recordings of their faces and free-text descriptions of their memories.

6.3.1. DATA COLLECTION PROCEDURE

We collected data from 300 crowd-workers via *Amazon Mechanical Turk*, providing a compensation of 6 *USD* each. Before any data collection, crowd-workers had to give their informed consent regarding the study procedure and all aspects of data collection and future use. Subjects first filled in a survey with additional information about themselves and their current situation. Then, we exposed each to a random selection of 7 stimuli from our pool of 42 music videos (see below for the selection of stimuli). During the playback, we recorded the participants' faces with their device (*Face Recordings*). After each clip, we requested ratings for the emotions it had induced (*Induced Emotion*), followed by a questionnaire about whether the video had caused them to recollect any personal memories. If this was the case, subjects were required to describe these memories with

Table 6.1: Response Data Overview

		$M\ (SD)$	Min/Max
Induced Emotion	**Pleasure**	0.29 (0.53)	-1.00/1.00
$N = 932$	**Arousal**	-0.03 (0.80)	-1.00/1.00
	Dominance	0.25 (0.58)	-1.00/1.00
Mem.-Assoc. Affect	**Pleasure**	0.34 (0.53)	-1.00/1.00
$N = 932$	**Arousal**	0.05 (0.79)	-1.00/1.00
	Dominance	0.30 (0.58)	-1.00/1.00
Memory Descr.	**Word No.**	25.07 (15.45)	3/103
$N = 932$			
Face Recordings	**Length (s)**	60.44 (2.10)	50.33/69.27
$N = 932$	**Frame No.**	1812.87 (63.28)	1510/2078

a short text (*Memory Descriptions*) and additional ratings of their feelings about them (*Memory-Associated Affect*). This procedure resulted in a total of 2098 unique responses from the participants. Out of these, a total of 978 responses of 260 unique participants included the recollection of at least one memory. We focus only on the subset of these responses for our experiments, for which also viable face recordings exist. After filtering out corrupted cases (e.g. malformed video data or incomplete recordings), this resulted in a combined set of 932 responses (see *Table 6.1*).

6.3.2. VIDEO STIMULI

We collect responses from viewers to a selection of 42 music video segments from among a set of 150 that were previously evaluated for their induced affect as part of creating the DEAP dataset [34]. We chose these stimuli for two reasons: 1. the strong capacity of music to trigger personal memories [22], and 2. existing PAD ratings from multiple viewers for each evaluated video. We hypothesized that responses to stimuli with low variation across viewers' PAD-ratings might be more directly driven by video content, and as such, either not produce or not be influenced by sources of person-specific variation, such as personal memories. For this reason, we used the existing ratings to balance our selections videos for low and high variation responses.

6.3.3. RESPONSE DATA

Induced Emotions: We asked participants to provide self-reports on their emotional responses to videos as pleasure-, arousal- and dominance-ratings. For this, they rated their experiences with the *AffectButton* instrument on a continuous scale in the interval of $[-1,+1]$. This rating tool is a 2d-widget displaying an iconic facial expression that changes in response to users' mouse or touch interactions. They can then provide ratings by selecting the facial expression that best fits the affect they want to express (see [41] for a detailed description and a validation study).

Memory Descriptions: Memory descriptions had to be provided in English and contain a minimum of three words. For each video, subjects could report as many memories as they had experienced. However, only 51 out of the 978 responses for which videos had triggered any memories involved 2 or more. For such multi-memory cases, we use the PAD ratings for memory-associated affect to identify the single memory in the response with the highest intensity of affect and retain only this in the modeling dataset. This filtering resulted in a total of 978 memory descriptions – one for each viewers' response.

Face Recordings: Recordings were captured by the devices that participants used when engaging with our online data collection application in their browser. While we enforced some constraints (e.g., to perform the task in a quiet setting), recordings are captured in conditions that are largely uncontrolled, reflecting the diverse ways in which people engage with media content in their daily lives. Therefore, recordings possess a wide range of different lighting conditions, are captured with different quality devices, and show crowd-workers changing postures (and even places). We transcoded all recordings from their original format to 30 frames per second. Several collected clips were corrupted by showing only a black screen, containing multiple individuals or encoding errors. Moreover, some possessed a duration abnormally shorter or longer than the 60 seconds of our video clips. We retained only uncorrupted recordings in the range of 50-70 seconds for the modeling activities reported in this chapter. This filtering left us with a set of 932 recordings of viewers' responses for which both memory descriptions and behavior are available.

6.4. PREDICTIVE MODELING

6.4.1. OVERVIEW

In line with most previous work on affect detection using dimensional representations, we address modeling viewers' emotional responses as a regression problem [3]. Support Vector Machines are a widely deployed approach when modeling affective responses to media content, especially in regression settings (see the reviews of technical work by Wang et al. [17], and more recently Zhao et al. [18]). For this reason, we use Support Vector Regressors with a Radial Basis Function (RBF)-kernel as predictors in our experiments.

An essential aspect of building context-sensitive affect detection is how information from different modalities is integrated into a single prediction, i.e., multimodal fusion. Existing work has primarily relied on either feature- or decision-level fusion of modalities, with neither approach showing clear superiority over the other [3]. However, previous work in which we explore both types of fusion for video content and memory-descriptions indicates a stronger overall performance of a decision-level approach using stacked generalization on this task, compared to feature-level fusion [31]. Motivated by this, we conduct all our experiments using only this approach to decision-level fusion. *Figure 6.1* provides a graphical overview of the entire machine learning pipeline that we deploy for predictions of induced emotions. Processing is undertaken in a traditional two-stage approach of *feature extraction* and *multimodal prediction*. The pipeline is deployed separately for predicting pleasure, arousal, and dominance.

An overview of the different information sources that we use as inputs and the feature-

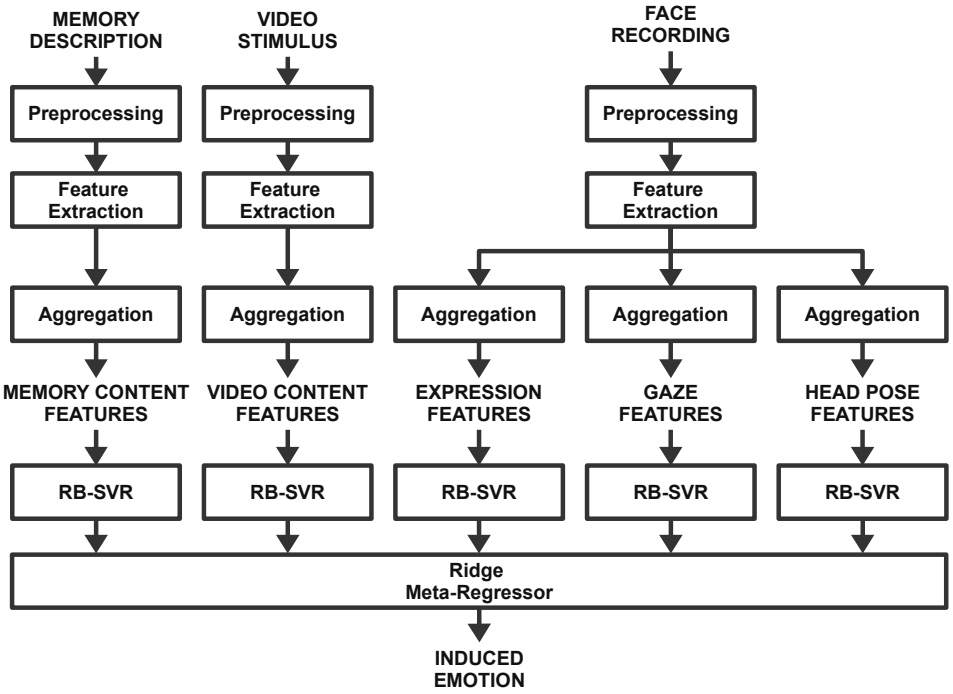


Figure 6.1: Overview of our approach for predictive modeling and decision-level multimodal fusion.

Table 6.2: Overview of Extracted and Modality-specific Feature Sets from Input Sources and their Aggregation

MODALITY	FEATURES	# EXTRACTED	SOURCE	# AGGREGATED
E	Action Units	17	Face Recording	13498
G	Direction	8	Face Recording	6352
P	Position/Orientation	6	Face Recording	4764
V	Theory-inspired	271	Video Stimulus	271
	Deep Visual	4096		4096
	Visual Sentiment	4342		4342
	openSMILE	1582		1582
M	Lexical	130	Memory Description.	130
	Word Embeddings	500		500

E: Facial Expressions; *G*: Gaze; *P*: Head Pose; *V*: Video Content; *M*: Memory Content

sets that we extract from them comprising different modalities for fusion and predictions can be found in *Table 6.2*. In total, we extract features from three different input sources: 1. recordings of viewers' faces, 2. the video stimuli that they are exposed to, and 3. free-text descriptions of triggered memories. The outcome of preprocessing and feature extraction in the first stage are 5 distinct feature-sets denoting different modalities for predicting viewers' response: 1. Facial Expressions, 2. Gaze, 3. Head Posture, 4. Video Content, and 5. Memory Content. Details about the preprocessing and feature extraction stages for each of these modalities are listed below. We extract many of these modality features from the input sources on a per-frame- or per-word-basis. For predictions, we aggregate these to the response level using statistical functions. Note that the extraction and aggregation stages for video stimuli and memories are identical to those described in our earlier work [31]. In the second stage, each aggregated modality-specific feature set is provided as input into a Support Vector Regressor for predictions. Finally, we fuse the outcome of these modality-specific models at the decision-level via stacking by an L2-regularized linear model ("Ridge" regression). All machine learning models use the implementation from the python library *Scikit-Learn* [42].

6.4.2. FACE RECORDINGS PROCESSING

We deploy the software *OpenFace 2.0* [43] for extracting feature-sets for *Expressions*, *Head Pose*, and *Gaze* from the face recordings in our dataset at the level of individual frames. All frame-level features for a recording are concatenated along the time-axis, and each resulting time series is aggregated to the response-level using statistical functions. For this purpose we rely on the *tsfresh* python package [44], which implements 63 best practice methods for time series characterization, computing a total of 794 generic features [1] per series. See *Table 6.2* for details about the amount of extracted and aggregated features per response.

Facial Expressions: OpenFace extracts information about facial muscle movements and expressions in terms of a subset of the *Facial Action Coding System (FACS)*. This coding scheme allows fine-grained descriptions of complex facial configurations by decomposing them into the activation of the combination of 45 individual muscles, i.e., Action Units. It is a widely used scheme for the objective characterization of facial expressions. For our model, we extract the intensity of activation of the 17 Facial Action Units provided by OpenFace (*AU Intensities*). Intensities range from $0-5$, whereby a value of 0 denotes no activation of the action unit in question, and a value of 5 an activation at maximum intensity. We drop any frames in videos with corrupted predictions (i.e., that are non-numeric or fall outside the $0-5$ range specified by the OpenFace developers for valid AU intensities). This filtering resulted in the exclusion of 3886 frames.

Gaze: In addition to facial expressions, we extract features about viewers' gaze direction as a distinct modality for predictions from each frame. They consist of an 8-dimensional vector, containing the (X, Z, Y) gaze *direction* in world coordinates for each eye separately and the horizontal and vertical gaze angles.

[1] A detailed list of the types of extracted time-series features is available here: https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html

Head Pose: Finally, we extract features describing the *location* and *orientation* of a person's head in relation to the camera to capture head pose as a distinct modality for predictions. Location is provided as a three-dimensional vector by denoting the (X, Y, Z)-position of the head in millimeters relation to the camera. On the other hand, orientation information is a vector of radians marking the pitch, yaw, and roll around the camera. Together, this results in the extraction of a 6-dimensional feature vector.

6.4.3. VIDEO STIMULUS PROCESSING

For the representation of the content of video stimuli as a modality in prediction, we extract different features from their visual and audio-tracks (see below). For visual analysis, we first export one frame per second of the video and extract features from it. The resulting frame-level feature vectors are then concatenated along the time axis, and aggregated by taking the mean. For extracting audio-features, we first split each video's audio track into a separate file, before using an existing software solution for processing (*openSMILE*). This software provides aggregated feature vectors of a fixed length to characterize the entire audio signal. See below for details about the extracted audio and visual features.

Theory-inspired Descriptors: Research on affective visual content analysis has developed descriptors inspired by psychology and art theory. We use a set of such descriptors developed by Machajdik & Hanburry [45], as well as those of Bhattacharya et al. [46] to characterize each of the extracted video frames. This combination has been used previously in affective content analysis (e.g., [47]).

Deep Visual Descriptors: Deep learning forms an essential part of the automatic analysis of image data. Instead of relying on engineered visual input descriptors, deep models can learn effective and reusable representations for prediction tasks from training data. We use the activation of the FC1-layer of a pre-trained VGG16 network [48] from the Keras framework for python [49] as features to capture a video frame's visual content (4096 dimensions). This representation has been used extensively as a baseline in benchmarking challenges for affective content analysis [50].

Visual Sentiment Descriptors: Prior research has established automatic detections of *Adjective-Noun Pairs (ANPs)* in visual material as useful high-level features for describing the affective content of visual stimuli (e.g., [5, 47]). ANPs are labels that denote objects or persons in an image, coupled with an affective attribute (in the spirit of "creepy forest"). We use the class-probabilities assigned by the *DeepSentiBank* Network [51] for any of the ANPs in its ontology as features describing a frame's content.

openSMILE: To represent the audio content of the music videos in our dataset we rely on the software *openSMILE* in the configuration *"emobase2010"* for feature extraction. It derives low-level descriptors from audio signals in a windowed fashion and aggregates them statistically into a single feature vector (see [36] for a detailed description). Benchmarking challenges for affective content analysis have used these features as a baseline approach [50].

6.4.4. MEMORY DESCRIPTIONS PROCESSING

We first clean memory descriptions by replacing references to specific years or decades (e.g., "1990", or "the 90s") with generic terms (e.g. "that year" or "that decade"). Additionally, we replace any numbers with 0 and expand all contractions present (e.g., "can't" is transformed into "cannot"). To model the affective impact of personal memories we extract word-level features that have proven successful in state-of-the-art models for predicting emotional states from social media text in a regression setting (see [29]): 1. *Lexical Features* and 2. *Word Embeddings* (see below for details). We then concatenate all word-level features in order of their appearance in the description, before taking the average to create a description-level representation.

Lexical Features: These features are created by parsing descriptions into word-level tokens and retrieving associated affective ratings from various affective dictionaries. We apply lemmatization before the lookup to remove word inflections to account for differences between words in descriptions and the form contained in lexica. The combination of the dictionaries that we initially selected for feature extraction [52–62] has achieved state-of-the-art performance for affect regression [63]. We extended this list by a new source containing word-level ratings for Pleasure, Arousal, and Dominance [29], and lexica-based VADER Sentiment ratings [64]. We aggregate word-level ratings to the description-level by averaging.

Word Embeddings: We leverage two pre-trained word embedding-models to represent each word in the memory description texts as a real-valued feature vector: *(1) Word2Vec*-model pre-trained on the *Google News dataset*, resulting in a 300-dimensional feature vector when applied to a word, and *(2)* a *GloVE*-model [65] pre-trained on the *Wikipedia 2014 and Gigaword 5 corpora*. It encodes individual words as a 200-dimensional feature vector. For both implementations we rely on the *Gensim*-library for python [66].

6.5. EMPIRICAL INVESTIGATION

To explore the influence of memory and video-content as contextual information for facial behavior in predictions, we conduct an ablation study of our model. This approach exhaustively compares the relative contributions of each modality and their multimodal combinations when predicting video-induced pleasure, arousal, and dominance. Notably, we collect samples for the test-performance of our model when having access to different modalities and conduct statistical analyses to quantify the contributions of context modalities 1. across affective dimensions (i.e., do they improve our model's overall performance?), as well as 2. within specific dimensions (i.e., do they provide our model with insights into some particular aspects of viewers' experience?).

6.5.1. EXPERIMENTAL SETUP

For training and evaluation of our model, we rely on nested 5-Fold-Leave-Persons-Out Cross-Validation. This procedure creates folds in such a way that no data from the same person is simultaneously available for both training and evaluation. The outer loop of the nested cross-validation splits the entire dataset into 5 folds, from which we hold out

a single fold for testing the performance of selected models. The inner loop uses the remaining 4 folds for optimizing the hyperparameters of the machine learning models through a grid search. To gain a better estimate of the influence of different modalities on the test performance of models, we repeat this procedure 6-times, resulting in samples of $N = 30$ data points of test performance for each investigated combination of modalities.

6.5.2. RESULTS AND ANALYSIS

A graphical overview of the distribution of test performance (R^2_{Test}) achieved by our model when provided with access to different combinations of modalities can be seen in *Figure 6.3*. Furthermore, *Table 6.2* provides the results of a statistical analysis of the differences between these samples of test performance[2].

Comparisons of Unimodal Performance vs. None-Baseline: We assess whether and which individual modalities facilitate an average test performance (R^2_{Test}) that is significantly above a baseline that always predicts the sample mean of the target variable in the development set used to build it ($None$). One-sided t-tests of performance samples where our model has only access to gaze- or head posture modalities indicate no improvement over this baseline. For this reason, we exclude them from all further analyses. Moreover, a look at the performance of modalities across targeted affective dimensions shows that memory content offers the highest individual performance for pleasure. In contrast, for predicting arousal, facial expressions provide the best performance, while the best performing modality for dominance is video content. This spread is an indicator of the overall complementary nature of these modalities for predictions of induced emotions.

Comparison of Multimodal Performance vs. Facial Expressions: In addition to individual modalities' performance, we tested whether combinations of context information and facial expressions result in improved model performance. For this purpose, we conduct paired t-tests between performance samples from models using only facial expressions (E) with those from having additional access to video content (V) or memory descriptions (M). We also compare the performance of having only access to both context sources ($V + M$) to facial expressions. These comparisons reveal that analyzing memory descriptions and video content provides substantial benefits to facial analysis for predicting pleasure, arousal and dominance.

Relationship between Modalities and Test-Performance: To further understand the relationship between our model's access to individual modalities and its test-performance within and across affective dimensions, we conduct a multi-way analysis of variance. For this purpose, we construct a linear mixed-effects model with R^2_{Test} as the dependent variable. We include fixed-effects for 1. the type of affective dimension targeted by the model (DIM), 2. access to Facial Expressions (E), 3. Video Content (V), 4. Memory

[2] Because we have obtained samples for test-performance R^2_{Test} from different repetitions of the nested cross-validation scheme these are no longer independent. However, following procedures outlined by Field et al. [67] we assessed the need for a hierarchical analysis using linear mixed-effects models to account for this nesting in comparisons within affective dimensions and found no significant improvements over simple linear models. Consequently, we stick to the more common procedures for statistical analysis resulting in Table 6.2.

Figure 6.2: Comparison of the test-performance of our model (R^2_{Test}) when predicting Induced Pleasure, Arousal and Dominance with access to only individual (vs. *None*-baseline) or multiple modalities (vs. only Facial Expressions (E)

	INDUCED PLEASURE				INDUCED AROUSAL				INDUCED DOMINANCE			
	R^2_{Test}	vs *None*			R^2_{Test}	vs *None*			R^2_{Test}	vs *None*		
Unimodal	*M (SD)*	$\Delta R^2{}_{Test}$	$t(df)$	p	*M (SD)*	$\Delta R^2{}_{Test}$	$t(df)$	p	*M (SD)*	$\Delta R^2{}_{Test}$	$t(df)$	p
None	-0.01 (0.01)	–	–	–	-0.01 (0.01)	–	–	–	–	–	–	–
E	0.02 (0.03)	0.03	5.96 (29)	<.001***	0.04 (0.04)	0.05	8.15 (29)	<.001***	0.03 (0.03)	0.04	8.97 (29)	<.001***
G	-0.01 (0.02)	-0.00	-1.13 (29)	.87	-0.01 (0.02)	-0.00	-1.27 (29)	.89	-0.01 (0.01)	-0.00	-1.04 (29)	.85
P	-0.01 (0.01)	-0.00	-1.3 (29)	.9	-0.01 (0.02)	0.00	0.04 (29)	.48	-0.01 (0.02)	0.00	1.58 (29)	.06
V	0.11 (0.04)	0.12	18.85 (29)	<.001***	0.02 (0.02)	0.03	9.52 (29)	<.001***	0.08 (0.03)	0.09	14.43 (29)	<.001***
M	0.15 (0.05)	0.16	17.13 (29)	<.001***	0.02 (0.03)	0.03	6.19 (29)	<.001***	0.06 (0.03)	0.07	14.21 (29)	<.001***
	R^2_{Test}	vs E			R^2_{Test}	vs E			R^2_{Test}	vs E		
Multimodal	*M (SD)*	$\Delta R^2{}_{Test}$	$t(df)$	p	*M (SD)*	$\Delta R^2{}_{Test}$	$t(df)$	p	*M (SD)*	$\Delta R^2{}_{Test}$	$t(df)$	p
$E+V$	0.12 (0.04)	0.10	20.02 (29)	<.001***	0.06 (0.04)	0.02	8.25 (29)	<.001***	0.1 (0.03)	0.07	11.49 (29)	<.001***
$E+M$	0.16 (0.06)	0.14	16.12 (29)	<.001***	0.06 (0.03)	0.02	5.73 (29)	<.001***	0.09 (0.04)	0.06	12.67 (29)	<.001***
$V+M$	0.19 (0.05)	0.17	20.1 (29)	<.001***	0.04 (0.03)	0.00	0.24 (29)	.41	0.11 (0.04)	0.08	10.15 (29)	<.001***
$E+V+M$	0.2 (0.05)	0.18	22.97 (29)	<.001***	0.07 (0.04)	0.04	8.41 (29)	<.001***	0.13 (0.04)	0.09	16.55 (29)	<.001***

None: Predictions use mean of target in development-set; E: Facial Expressions; G: Gaze; P: Head Pose; V: Video Content; M: Memory Content;

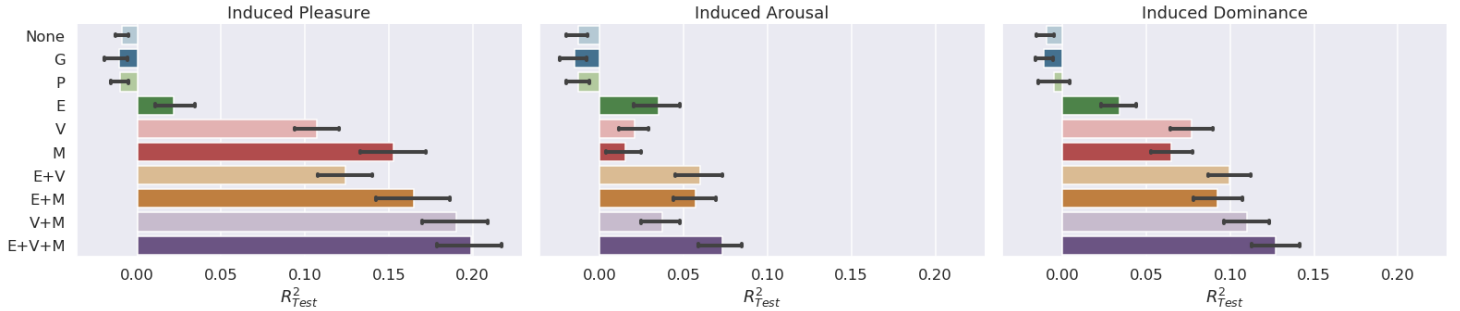


Figure 6.3: Test-performance (R^2_{Test}) of our model for Induced Pleasure, Arousal and Dominance when using Facial Expressions (E), Gaze (G), Head Pose (P), Video Content (V), Memory Content (M), or their multimodal fusions for predictions. *None* is a baseline always predicting the mean of targets in the development-set for tests. Error bars denote the 95% confidence interval.

Table 6.3: Effects of Modalities and Targeted Affect Dimension on Model Performance (R^2_{Test})

Effect	df_n	df_d	F	p
E	1	701.999	120.738	<.001***
V	1	701.999	388.018	<.001***
M	1	701.999	550.757	<.001***
DIM	2	18.000	263.067	<.001***
$E*V$	1	701.999	4.124	<.05*
$E*M$	1	701.999	3.333	.068
$V*M$	1	701.999	58.049	<.001***
$E*DIM$	2	701.999	7.194	.01**
$V*DIM$	2	701.999	30.274	<.001***
$M*DIM$	2	701.999	117.646	<.001***
$E*V*M$	1	701.999	0.563	.454
$E*V*DIM$	2	701.999	0.254	.776
$E*M*DIM$	2	701.999	0.243	.784
$V*M*DIM$	2	701.999	12.374	<.001***
$E*V*M*DIM$	2	701.999	0.070	.932

DIM: Targeted Affect Dimension; E: Facial Expressions; V: Video Content; M: Memory Content;

Content (M) modalities, as well as 5. their multi-way interactions. To account for the nesting of samples in our analysis, we include random effects dependent on the identity of repetitions (maximum random effects structure supported by the data is determined empirically; resulted in intercept only).

As expected, the results of this analysis in *Table 6.3* show significant main-effects for each modality on model performance. The positive coefficients of these effects indicate that access to each modality has a significantly positive impact on performance across affective dimensions (E: $b = 0.05$; V: $b = .03$; M: $b = 0.02$). Moreover, average test performance is greater when models have access to memory content compared to video content (Mvs$V$: $t(29) = 2.17, p < .05$), or facial expressions (Mvs$E$: $t(29) = 9.37$, $p < .001$). Apart from this, there is a significant effect of DIM on test performance, showing that our model's average performance varies systematically across affective dimensions, independent of the modalities involved.

Further, inspection reveals no significant interactions between the context modalities and facial expressions ($E*V*M$ or $E*V*M*DIM$), indicating that – independent of the targeted affective dimension – no substantial overlap in provided information exists between them. This finding demonstrates the complementary nature of context information for facial analysis. In contrast, there is a significant interaction between memory- and video-content ($V*M$), indicating overlap. The coefficient for this effect in the analysis reveals the negative influence of this interaction on model performance ($b = -0.01$), showing that their benefits diminish when both modalities are accessible. Moreover, this interaction's strength seems to depend on the affective dimensions targeted by models ($V*M*DIM$). A glimpse at the interaction plots in *Figure 6.4* provides further insights into the nature of this relationship. Especially when predicting pleasure, video, and memory content provide overlapping information for our model, reducing their positive impact on performance.

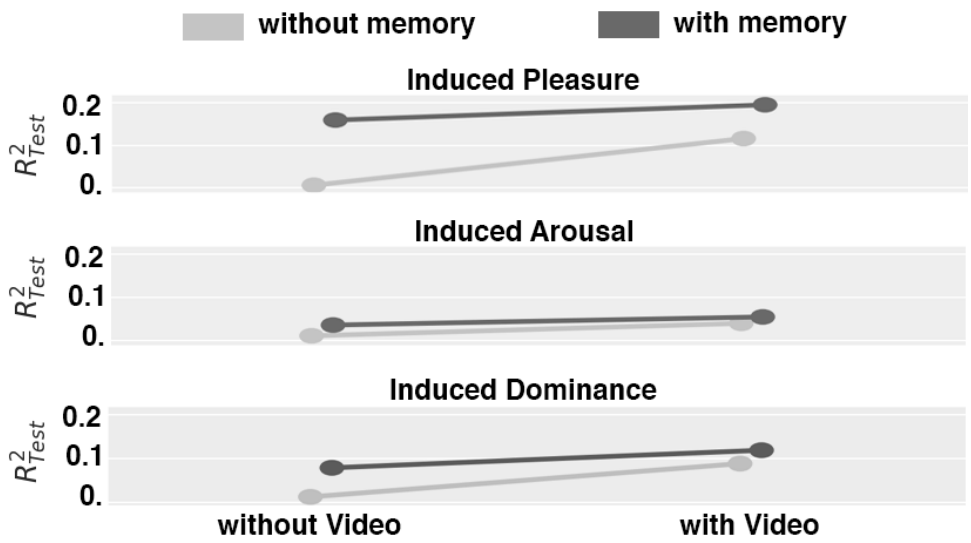


Figure 6.4: Marginal means of Model Performance (R^2_{Test}) with/without access to Memory and Video Content modalities. Converging lines indicate negative interactions due to overlapping information.

6.6. DISCUSSION

6.6.1. EMPIRICAL FINDINGS

The findings from our empirical investigation demonstrate that information about what viewers are watching, and what that reminds them off is is highly complementary to the insights offered by analysis of their facial behavior. The benefits of the video- and memory-content modalities for predictions manifest both by increasing the average performance of models across affective dimensions and offering specific benefits for individual affective dimensions. Depending on what aspect of affective experience applications are interested in, they may benefit from knowledge about some contextual influence more than knowledge about others. Furthermore, our results indicate that viewers' self-reported memory descriptions provide substantial insights across affective dimensions in our experiments. This finding is congruent with our earlier investigations, where we compared the performance of memory descriptions for predictions to that of only video content [31]. This capacity of text-based memory descriptions for predicting emotional responses should motivate computational research to provide automatic systems with access to this information. The first step towards this could be to mine video-associated memory descriptions from social media content, e.g., by automatically identifying relevant user comments. Moreover, technological approaches could explore how the emotional meaning of already collected memory descriptions relates to novel viewing situations and videos. More generally, personal memories form a crucial contextual driver for video-induced emotions [23], and accounting for their impact in automatic predictions could facilitate a broad range of novel applications [68], e.g., affect-based reminiscence support technology. More comprehensively addressing personal memories forms a substantial challenge for computational modeling because of their person- and

situation-specific nature. Doing so requires – apart from technological contributions – also developing datasets and corpora that capture the occurrence and emotional impact of memories on responses.

Apart from insights about the context, our results indicate that the affective information provided by facial behavior in isolation is comparatively low. This observation is congruent with the findings of Hirt et al. [9], demonstrating an overall lack of correspondence between face-based affect predictions and emotional experience in a human-computer interaction setting. One possible explanation is that people scarcely express their emotions through the face when viewing videos alone on their devices. Psychological theory overall argues for the essential social functions of emotional expressions [69], e.g., facilitating bonds with others. As such, there may be little functional need for displaying them in single-person settings. If this is the case, the usefulness of facial behavior for predictions in such a setting may be inherently limited. However, it is important to note that our analyses of facial behavior rely on data automatically extracted through OpenFace. The automatic analysis of the face recordings in our dataset is a substantial challenge for existing technology: lighting conditions vary, viewers move or change position, etc. These adverse conditions likely hurt the accuracy with which the OpenFace-software can extract facial features, providing an alternative explanation for their relatively low value for predictions. Ultimately, however, our current study cannot differentiate with certainty whether participants' low expressivity or error in automatic recognition is the cause for the relatively low performance of predictions based on facial behavior. However, analysis of facial expressions consistently facilitates performance across all dimensions of viewers' affective states, outperforming both context-modalities for arousal predictions. This finding further highlights the necessity of combining different information sources in automatic affect detection to achieve accuracy and robustness in-the-wild.

6.6.2. LIMITATIONS

Despite the insights provided by our empirical investigation, there are several methodological limitations to their validity. For once, an additional explanation for our model's comparatively weak performance when relying on facial behavior might be that it fails to exploit the rich temporal context of these behavioral signals sufficiently. More sophisticated temporal modeling techniques explored in affect recognition, e.g., LSTMs, might result in better absolute performance, but also require large corpora for training [32]. Another explicit limitation of our approach is that we analyze only responses in our dataset for which viewers reported having recollected memories. However, the information provided by facial behavior about emotional experiences may differ when no memories are involved. For example, gaze patterns might provide more information in this case, because visual content is more directly driving responses. Future research could explore such differences in facial behavior patterns during video-induced emotions more directly. Finally, while we explicitly instructed participants only to report memories if they had experienced them *during* the video, the sequence in which we asked for affective self-reports may affect whether and what memories are recollected, or how they are evaluated. Future investigations should actively minimize such influences in their study design, e.g., by spacing out describing and evaluating memory content over time.

6.7. SUMMARY AND CONCLUSION

Analysis of individuals' facial behavior is an extensively researched approach for automatic detection of affect. However, the emotional meaning of facial expressions in isolation can be ambiguous. For this reason, humans extensively rely on potential causes for the emotions experienced by others as additional context for their inferences. Apart from videos' content, an essential cause for emotional responses is the triggering of viewers' personal memories. This chapter has explored the impact of providing an automatic affect detection system with additional information about both of these two influences to contextualize the analysis of viewers' facial behavior. Our machine learning experiments' findings indicate that this combination facilitates more accurate predictions than looking at facial behavior in isolation. Moreover, while adding context information improves models' overall accuracy, individual sources provide particular advantages for predicting specific affective dimensions. This complementary nature of sources means that application developers might make meaningful trade-offs by choosing which information to incorporate for predictions. More generally, awareness of contextual influences may facilitate more accurate predictions and provide clear and immediate benefits for downstream tasks to build on them meaningfully (e.g., by reacting adequately to the likely cause of viewers' emotional response). Predicting emotions in a video-viewing setting may be particularly suitable for exploring aspects of context and their integration into affect detection because it is relatively clearly defined and constrained regarding potential influences compared to other types of situations.

Overall, our investigations reveal the analysis of viewers' memory descriptions as a substantial source of information about their affective responses. For this reason, affect-detection systems can benefit from technological research that provides them as input for predictions, e.g., by automatically mining memory descriptions from viewers' social media comments or associating existing memory descriptions with new video content. Ultimately, however, only computational modeling that systematically explores predicting occurrence (when?), content (what?), and influence (what does it do?) can adequately address the influence of personal memories as a context for predictions emotional responses.

REFERENCES

[1] A. Bartsch, *Emotional Gratification in Entertainment Experience. Why Viewers of Movies and Television Series Find it Rewarding to Experience Emotions,* Media Psychology **15**, 267 (2012).

[2] A. Hanjalic and Li-Qun Xu, *Affective video content representation and modeling,* IEEE Transactions on Multimedia **7**, 143 (2005).

[3] S. K. D'mello and J. Kory, *A Review and Meta-Analysis of Multimodal Affect Detection Systems,* ACM Computing Surveys **47**, 1 (2015).

[4] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *A multimodal database for affect recognition and implicit tagging,* IEEE Transactions on Affective Computing **3**, 42 (2012).

[5] D. McDuff and M. Soleymani, *Large-scale Affective Content Analysis: Combining Media Content Features and Facial Reactions,* in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (IEEE, 2017) pp. 339–345.

[6] J. J. Sun, T. Liu, A. S. Cowen, F. Schroff, H. Adam, and G. Prasad, *EEV Dataset: Predicting Expressions Evoked by Diverse Videos,* (2020), arXiv:2001.05488 .

[7] J. M. Fernández-Dols and C. Crivelli, *Emotion and expression: Naturalistic studies,* Emotion Review **5**, 24 (2013).

[8] 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,* Psychological Science in the Public Interest **20**, 1 (2019).

[9] F. Hirt, E. Werlen, I. Moser, and P. Bergamin, *Measuring emotions during learning: lack of coherence between automated facial emotion recognition and emotional experience,* Open Computer Science **9**, 308 (2019).

[10] D. Dupré, E. G. Krumhuber, D. Küster, and G. J. McKeown, *A performance comparison of eight commercially available automatic classifiers for facial affect recognition,* PLOS ONE **15**, e0231968 (2020).

[11] U. Hess and S. Hareli, *The influence of context on emotion recognition in humans,* in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (IEEE, 2015) pp. 1–6.

[12] M. J. Wieser and T. Brosch, *Faces in context: A review and systematization of contextual influences on affective face processing,* Frontiers in Psychology **3**, 471 (2012).

[13] D. Matsumoto and H. Sung Hwang, *Judging Faces in Context,* Social and Personality Psychology Compass **4**, 393 (2010).

[14] A. Marpaung and A. Gonzalez, *Can an affect-sensitive system afford to be context independent?* in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10257 LNAI (Springer, Cham, 2017) pp. 454–467.

[15] B. Dudzik, M.-P. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong, *Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases,* in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 206–212.

[16] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, *Corpus Development for Affective Video Indexing,* IEEE Transactions on Multimedia **16**, 1075 (2014), arXiv:1211.5492 .

[17] S. Wang and Q. Ji, *Video Affective Content Analysis: A Survey of State-of-the-Art Methods,* IEEE Transactions on Affective Computing **6**, 410 (2015).

[18] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, *Affective Computing for Large-scale Heterogeneous Multimedia Data,* ACM Transactions on Multimedia Computing, Communications, and Applications **15**, 1 (2020), arXiv:1911.05609 .

[19] Z. Hammal and M. T. Suarez, *Towards context based affective computing introduction to the third international CBAR 2015 workshop,* in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (IEEE, 2015) pp. 1–2.

[20] D. G. McDonald, M. A. Sarge, S.-F. Lin, J. G. Collier, and B. Potocki, *A Role for the Self: Media Content as Triggers for Involuntary Autobiographical Memories,* Communication Research **42**, 3 (2015).

[21] A. M. Belfi, B. Karlan, and D. Tranel, *Music evokes vivid autobiographical memories,* Memory **24**, 979 (2016).

[22] P. Janata, S. T. Tomic, and S. K. Rakowski, *Characterisation of music-evoked autobiographical memories,* Memory **15**, 845 (2007).

[23] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, *Investigating the Influence of Personal Memories on Video-Induced Emotions,* in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (ACM, New York, NY, USA, 2020) pp. 53–61.

[24] B. Rimé, S. Corsini, and G. Herbette, *Emotion, verbal expression, and the social sharing of emotion,* The verbal communication of emotions: Interdisciplinary perspectives , 185 (2002).

[25] W. R. Walker, J. J. Skowronski, J. A. Gibbons, R. J. Vogl, and T. D. Ritchie, *Why people rehearse their memories: Frequency of use and relations to the intensity of emotions associated with autobiographical memories,* Memory **17**, 760 (2009).

[26] S. Bluck, N. Alea, T. Habermas, and D. C. Rubin, *A TALE of Three Functions: The Self–Reported Uses of Autobiographical Memory,* Social Cognition **23**, 91 (2005).

[27] B. Caci, M. Cardaci, and S. Miceli, *Autobiographical memory, personality, and Facebook mementos,* Europe's Journal of Psychology **15**, 614 (2019).

[28] D. Cosley, V. S. Sosik, J. Schultz, S. T. Peesapati, S. Lee, T. Peesapati, and S. Lee, *Experiences With Designing Tools for Everyday Reminiscing,* Human–Computer Interaction **Volume 27**, 175 (2012).

[29] S. Mohammad, *Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words,* in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018) pp. 174–184.

[30] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, *SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text,* in *Proceedings of the 13th International Workshop on Semantic Evaluation* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019) pp. 39–48.

[31] 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 .

[32] P. V. Rouast, M. Adam, and R. Chiong, *Deep Learning for Human Affect Recognition: Insights and New Developments,* IEEE Transactions on Affective Computing (2018), 10.1109/TAFFC.2018.2890471.

[33] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, *Emotion Recognition in Context,* in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017) pp. 1960–1968.

[34] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *DEAP: A Database for Emotion Analysis ;Using Physiological Signals,* IEEE Transactions on Affective Computing **3**, 18 (2012).

[35] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, *AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups,* Expert Systems with Applications **39**, 12378 (2017), arXiv:1702.02510 .

[36] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, *The INTERSPEECH 2010 paralinguistic challenge,* in *Eleventh Annual Conference of the International Speech Communication Association* (2010).

[37] A. Mehrabian, *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament,* Current Psychology **14**, 261 (1996).

[38] P. J. Lang, M. M. Bradley, B. N. Cuthbert, and Others, *International affective picture system (IAPS): Technical manual and affective ratings,* NIMH Center for the Study of Emotion and Attention **1**, 39 (1997).

[39] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, *EMOTIC: Emotions in Context Dataset,* in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Vol. 2017-July (2017) pp. 2309–2317.

[40] J. Broekens, *In Defense of Dominance,* International Journal of Synthetic Emotions **3**, 33 (2012).

[41] J. Broekens and W.-P. Brinkman, *AffectButton: A method for reliable and valid affective self-report,* International Journal of Human-Computer Studies **71**, 641 (2013).

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and Others, *Scikit-learn: Machine learning in Python,* Journal of machine learning research **12**, 2825 (2011).

[43] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, *OpenFace 2.0: Facial Behavior Analysis Toolkit,* in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (IEEE, 2018) pp. 59–66.

[44] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, *Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package),* Neurocomputing **307**, 72 (2018).

[45] J. Machajdik and A. Hanbury, *Affective image classification using features inspired by psychology and art theory,* in *Proceedings of the international conference on Multimedia - MM '10* (ACM Press, New York, New York, USA, 2010) p. 83.

[46] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, *Towards a comprehensive computational model foraesthetic assessment of videos,* in *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, 3 (ACM Press, New York, New York, USA, 2013) pp. 361–364.

[47] M. J. Scott, S. C. Guntuku, W. Lin, and G. Ghinea, *Do Personality and Culture Influence Perceived Video Quality and Enjoyment?* IEEE Transactions on Multimedia **18**, 1796 (2016).

[48] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition,* in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2015) arXiv:1409.1556 .

[49] F. Chollet and Others, *Keras,* \url{https://keras.io} (2015).

[50] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, Z. Xiao, and M. Sjöberg, *The MediaEval 2018 emotional impact of Movies task,* CEUR Workshop Proceedings , 1.

[51] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, *DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks,* (2014), arXiv:1410.8586 .

[52] M. Hu and B. Liu, *Mining and summarizing customer reviews,* in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04* (ACM Press, New York, New York, USA, 2004) p. 168.

[53] S. Baccianella, A. Esuli, and F. Sebastiani, *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.* in *Lrec*, Vol. 10 (2010) pp. 2200–2204.

[54] S. Mohammad and P. Turney, *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon,* in *Proceedings of the {NAACL} {HLT} 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (Association for Computational Linguistics, Los Angeles, CA, 2010) pp. 26–34.

[55] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, *Sentiment strength detection in short informal text,* Journal of the American Society for Information Science and Technology **61**, 2544 (2010).

[56] Y. R. Tausczik and J. W. Pennebaker, *The psychological meaning of words: LIWC and computerized text analysis methods,* (2010).

[57] F. Å. Nielsen, *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,* CEUR Workshop Proceedings **718**, 93 (2011), arXiv:1103.2903 .

[58] S. M. Mohammad, S. Kiritchenko, and X. Zhu, *NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets,* *SEM 2013 - 2nd Joint Conference on Lexical and Computational Semantics **2**, 321 (2013), arXiv:1308.6242 .

[59] Y. Choi and J. Wiebe, *+/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference,* in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014) pp. 1181–1191.

[60] S. M. Mohammad and S. Kiritchenko, *Using Hashtags to Capture Fine Emotion Categories from Tweets,* Computational Intelligence **31**, 301 (2015).

[61] F. Bravo-Marquez, E. Frank, S. M. Mohammad, and B. Pfahringer, *Determining Word-Emotion Associations from Tweets by Multi-label Classification,* in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (IEEE, 2016) pp. 536–539.

[62] S. M. Mohammad, *Word Affect Intensities,* LREC 2018 - 11th International Conference on Language Resources and Evaluation , 174 (2017), arXiv:1704.08798 .

[63] V. Duppada, R. Jain, and S. Hiray, *SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets,* Proceedings of The 12th International Workshop on Semantic Evaluation , 18 (2018), arXiv:1804.06137 .

[64] C. J. Hutto and E. Gilbert, *VADER: A parsimonious rule-based model for sentiment analysis of social media text,* in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (2014) pp. 216–225.

[65] J. Pennington, R. Socher, and C. Manning, *Glove: Global Vectors for Word Representation,* in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014) pp. 1532–1543.

[66] R. Řehůřek and P. Sojka, *Software Framework for Topic Modelling with Large Corpora,* in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA, Valletta, Malta, 2010) pp. 45–50.

[67] A. Field, J. Miles, and Z. Field, *Discovering statistics using R* (Sage publications, 2012).

[68] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, *Artificial Empathic Memory,* in *Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions - EE-USAD'18* (ACM Press, New York, New York, USA, 2018) pp. 1–8.

[69] A. H. Fischer and A. S. R. Manstead, *Social functions of emotion,* in *Handbook of emotions (3rd ed.)*, edited by M. Lewis, J. M. Haviland-Jones, and L. F. Barrett (Guilford Press, New York, NY, US, 2008) 3rd ed., Chap. 28, pp. 456–468.

7

SITUATING REMEMBERED EPISODES IN LIFELOG DATA

This chapter is based on: Dudzik, B., Olenick, J., Broekens, J., Chang, C. H., Hung, H., Neerincx, M., & Kozlowski, S. W. J. (2018). Discovering digital representations for remembered episodes from lifelog data. Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, MCPMD 2018.

ABSTRACT

Combining self-reports in which individuals reflect on their thoughts and feelings (Experience Samples) with sensor data collected via ubiquitous monitoring can provide researchers and applications with detailed insights about human behavior and psychology. However, meaningfully associating these two sources of data with each other is difficult: while it is natural for human beings to reflect on their experience in terms of remembered episodes, it is an open challenge to retrace this subjective organization in sensor data referencing objective time.

Lifelogging is a specific approach to the ubiquitous monitoring of individuals that can contribute to overcoming this recollection gap. It strives to create a comprehensive timeline of semantic annotations that reflect the impressions of the monitored person from his or her own subjective point-of-view.

In this chapter, we describe a novel approach for processing such lifelogs to situate remembered experiences in an objective timeline. It involves the computational modeling of individuals' memory processes to estimate segments within a lifelog acting as plausible digital representations for their recollections. We report about an empirical investigation in which we use our approach to discover plausible representations for remembered social interactions between participants in a longitudinal study. In particular, we describe an exploration of the behavior displayed by our model for memory processes in this setting. Finally, we explore the representations discovered for this study and discuss insights that might be gained from them.

7.1. INTRODUCTION

Experience Sampling Methods (ESMs) refer to a variety of approaches used by researchers for collecting self-reports (e.g. with questionnaires) from individuals about their subjective impressions, thoughts and feelings in the scope of their everyday lives [1]. Some studies have used these methods for the collection of data detailing subjects' experiences during specific situations, e.g. social interactions [2] or instances in which addicts experience craving [3].

Recently, studies have begun to combine this form of data collection with ubiquitous monitoring via wearable sensors, e.g. to investigate long-term team dynamics [4]. Such devices offer additional information about the behaviors displayed by participants, as well as their corresponding context. In combination, these two sources of information hold the potential to provide researchers with a detailed description of how complex social and psychological phenomena emerge and evolve over time [5].

However, an open challenge to unlocking the full potential offered by such a synchronized description is to unpack which sensor readings describe those moments in time that individuals are referring to in their self-reports. This is a difficult task, because of what we will refer to in the following as the *recollection gap*: in contrast to sensor data, the subjective impressions that people are sharing in this way are not referencing objective time periods. Instead, these are grounded in the recollections of their past as specific *Episodes*. These are mental constructs comprising slices of their previous experience. They are primarily defined in terms of their content (i.e. *"what"* they are about) [6], as well as their relative position within the remembering person's overarching life story [7].

While it is possible for people to provide an objective time for the episodes that they remember, it appears difficult for them to do so accurately or consistently (see e.g. [3]). Consequently, time-based information alone is of limited help in bridging this gap. Instead, we need to find a way of situating episodes within an objective timeline, based on those attributes that define them for the person undergoing recollection: elements of the episodic content experienced and associations with their personal history.

Lifelogging is a special approach to ubiquitous monitoring that can contribute towards such a human-centered approach for bridging the recollection gap. Instead of merely organizing data into a timeline, lifelogging provides automatically-generated semantic annotations along-side it. These are meant to approximate an individual's subjective impressions in the situations that he or she encounters while being monitored [8]. For example, a person may be equipped with a wearable camera whose recorded images are then automatically annotated with the labels of places or objects that are visible in them. Because these labels are based on data that was captured from a the subjective point-of-view of the person, they may act as meaningful proxies for the person's actual perceptions. To highlight this connection, we will explicitly refer to annotations created in such a fashion as *Perception Proxies*.

Importantly, these proxies may support anchoring remembered episodes in an objective timeline: the places, people, or objects that an individual experiences as part of an episode, may possess corresponding proxies within their collected lifelog timeline. Consequently, a segment of this timeline that corresponds with content of the recollected episode, may serve as a plausible representation for it. In essence, such *Digital Episode Representations (DERs)* allow an estimate of *when* a given episode may have occurred, and for *how long* it may have lasted.

In this chapter, we propose a novel approach for bridging the recollection gap by discovering such plausible representations for episodes of interest from lifelog data. In essence, it takes the form of a computational model of the memory processes that have resulted in the recollection of these specific episodes: when provided with a description of a target episode (an indication what was remembered by the person), it emulates the process leading to its recollection by extracting some segment from the lifelog (an indication of what has been experienced) that corresponds with it.

With respect to this, our primary contributions described in this chapter are the following:

- We present an approach for computationally modeling individuals' memory processes when responding to specific requests for information about their past.
- We give a detailed explanation of a computational model for the specific memory processes displayed by the participants in a longitudinal study, reflecting about social interactions with each other.
- We report on a series of empirical investigations in which we explore the behavior of our model for the recollections in this particular scenario, as well as the representations it discovers.

7.2. RELATED WORK

Important attributes that distinguish lifelogging from other approaches to the pervasive monitoring of individuals (such as surveillance) include: *1)* a focus on passive and continuous capture of data related to a *single individual* [8], *2)* the collection of data from a subjective point-of-view through wearable devices (e.g. [9]), and *3)* a focus on the automatic annotation of data-traces with labels that describe a person's subjective impressions (e.g. by naming places, objects or persons detected in visual data [10, 11]).

Technical approaches to construct lifelogs have adopted *events* as a basic unit of organization for timelines [8]. Different methods have been devised to provide automatic temporal segmentation of multimodal data streams in such a fashion (e.g. [12]). Similarly, research on lifelogging applications has explored the aggregation of semantic annotations in a timeline to provide relevant descriptions at this event-level [13]. However, the goal of such endeavors is not to discover representations for specific episodes . Rather, they try to create meaningful atomic units to manage and access the large collections of personal data that are being produced by lifelogging appliances [8, 14]. That is, their purpose is to facilitate generic information retrieval tasks. As far as we are aware, no other work in the lifelogging-domain has attempted to create digital representations for episodes in the sense that we describe here.

7.3. OUR APPROACH

In summary, the approach that we propose for discovering representations for remembered episodes consists of two steps:

1. Constructing a computational model for the specific memory processes that have lead to the recollection of the target episodes. In particular, this involves the specification of a process for evaluating a segment of lifelog data for its correspondence with these episodes.

2. Applying this model to lifelogs from the individuals that have remembered these target episodes, in order to identify plausible representations for them.

In *Section 4* we provide a general outline of our computational model for memory processes underlying the recollection of episodes when being asked for information about one's past. In *Section 5* we describe a dataset that was obtained as part of a longitudinal study, and contains information describing recollected episodes in addition to relevant lifelog timelines. It forms the context for an empirical investigation of our approach in *Section 6*. There we give an account of our computational model for participants' memory processes in this particular setting, and explore both its behavior its results when discovering DERs.

7.4. A COMPUTATIONAL MODEL OF MEMORY RESPONSES

Contemporary psychology generally agrees that access to memories describing personal experiences can take two basic forms: either they emerge on their own, based on associations with cues in one's environment, or one deliberately causes them by searching for information about the past [15]. Requesting someone to provide information about their

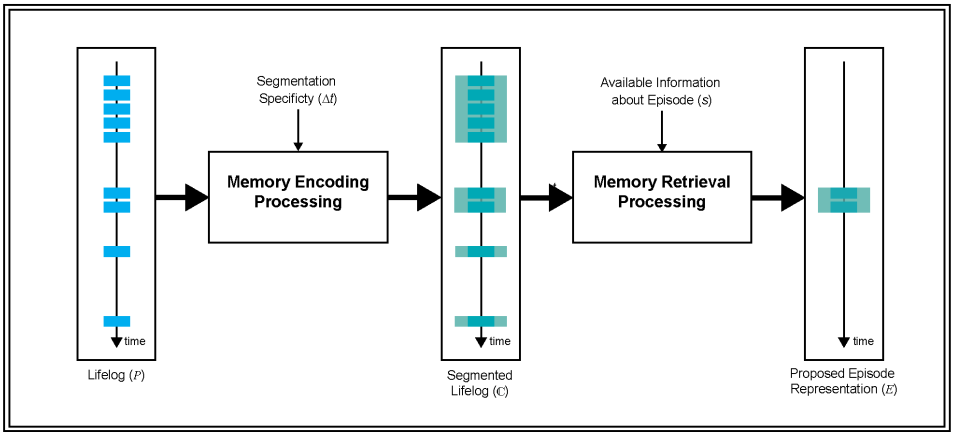


Figure 7.1: Overview of the proposed model for memory responses for discovering DERs. The *Memory Encoding Processing* splits a lifelog timeline into segments, which are evaluated for their correspondence with information about the episode at the stage of *Memory Retrieval Processing*. The segment with the greatest correspondence is proposed as a candidate for representation.

past, as is done in experience sampling, can be seen as instructing a person to initiate such a deliberate search. In essence, the precise instructions that a person is provided *define* some attributes that an episode needs to fulfill to be considered as relevant for recollection. In the following, we will summarily refer to all the cognitive processes that are undertaken by a person to answer such a request about their past as his or her *memory response*.

In this section, we introduce a computational model of such responses for the purpose of discovering DERs. We will provide a detailed description of the sub-processes constituting it, as well as the representations that it draws on (see *Figure 7.1* for an overview).

7.4.1. MEMORY ENCODING PROCESSING

Memory Encoding describes the cognitive process utilized by individuals to parse their continuous experiences into mental representations, which are later accessible as distinct episodes. An important principle in human cognition for integrating experienced stimuli into the same episodes is their consecutive temporal proximity to each other [16, 17]. That is, information that is experienced as occurring relatively close to each other, also tends to be recollected as part of the same episode.

The sub-process of *memory encoding processing* in our model operates according to this specific principle. It's purpose is to emulate the memory encoding that has preceded the recollection of a specific episode in a psychological plausible way.

When provided with a lifelog timeline, it splits it into a collection of non-overlapping segments by grouping temporally close perception proxies together. Each of these segments is then considered to be a potential candidate for representing the outcome of the modeled memory response, i.e. the episode for which a corresponding representation should be discovered.

For this purpose, let $P = \{p_0, p_1, \ldots, p_n\}$ denote a lifelog timeline wherein each element

is a timed perception proxy p. A perception proxy itself takes the form of a 3-tuple (t, a, o), were, t is a numerical timestamp that denotes when the entry has been created, a is a label that describes the content that it stands in for (e.g. the name of a specific place or object that was encountered by a person), and o is a unique identifier for the person from whose perspective it was created.

The process of memory encoding then is denoted by the function $enc(P, \Delta t)$. It partitions the contents of a lifelog timeline into a collection of non-overlapping segments $\mathbb{C} = \{C_0, C_1, \ldots, C_n\}$. This segmentation is regulated by the parameter Δt that denotes the amount of time that can pass between two consecutive perception proxies in the timeline P, before they are assigned to a different segment (*segmentation specificity*):

$$\forall C \in \mathbb{C}\Big(\forall i\Big(|t(p_{i+1}) - t(pi)| < \Delta t \wedge p_i \in C \wedge p_{i+1} \in C\Big)\Big) \tag{7.1}$$

7.4.2. MEMORY RETRIEVAL PROCESSING

This stage approximates those cognitive processes that have resulted in an individual's willful recollection of a specific episode as part of the modeled memory response. When provided with a collection of candidate segments formed from a given lifelog timeline, it assesses the degree to which each such segment corresponds with information that is available about the episode for which a DERs should be discovered.

To this end, it defines a computational evaluation procedure represented by some function $cor(C, s)$. Here $C \in \mathbb{C}$ is a specific candidate segment under evaluation, while s refers to a collection of available information about an episode that the individual has recollected as part of the modeled memory response. The computational procedure for this evaluation of each segment can take any information into account that is provided by the perception proxies in its timeline. The outcome is a numerical score in the interval $[0, 1]$. A result of 0 describes no correspondence with information describing the episode, while a 1 stands for the greatest possible degree of correspondence.

Given this, the lifelog segment that achieves the highest degree of correspondence is chosen as the most plausible candidate for representation of the episode:

$$E = \underset{C \in \mathbb{C}}{\operatorname{argmax}}\Big(cor(C, s)\Big) \tag{7.2}$$

7.5. THE DATASET

The dataset that we use for an empirical exploration of our approach to discover episode representations was collected as part of a longitudinal study about the dynamics of team-cohesion, and has been utilized in previously published work (e.g [18]). It describes the social interactions of six participants (here coded as *P1* to *P6*) within an isolated environment in the context of a simulated space mission.

For our purposes, two types of records that were collected are particularly relevant: *1)* a range of experience samples in which participants reflect about occurrences of social interactions with each other, and *2)* associated lifelog data from the perspective of each participant. In the following we will describe relevant aspects of these records and how they were collected in more detail. Because one of the participants (coded *P5*) withdrew

Table 7.1: Experience Samples per Participant.

	P1	P2	P3	P4	P6	Total
N	193	197	151	140	190	871

early from the study for personal reasons, we disregard those records entirely from both our description and modeling activities.

7.5.1. EXPERIENCE SAMPLES

Participants were instructed to provide structured reports about the occurrence of social interactions twice-daily at fixed times: once in the morning and once in the evening. Reports could be voluntarily provided at any time through a computer-based questionnaire. This questionnaire instructed participants to recollect and evaluate the most recent social interaction that they had engaged in with other members of the team. The information that they were required to provide about this interaction included the identity of their interaction partners. Moreover, each reported instance could also be annotated with one or more labels specifying the type of interaction it pertained to. Choices that participants were provided with included: *Task Interaction related to Team Goals (T)*, *Task Interaction related to Individual Responsibilities (I)* or *Social Interaction (S)*. Additional evaluations that were requested from them involved judgments of their experiences during the interaction, as well as its perceived effectiveness. Additionally, the time at which participants started and completed the form was automatically recorded by the system.

Table 7.1 lists the experience samples available for each participants.

A detailed look at this collection of reports also exemplifies some of the practical challenges of situating episodes within an objective timeline. While each experience sample possesses a timestamp for when itself was provided by a participant, this does not necessarily allow one to demarcate when the remembered episode itself took place. Especially problematic w.r.t. this is that participants appear to often cross the specified sampling intervals when providing their reports. This can be spotted in *Figure 7.2*: there is an over-proportionally large total share of samples present in the second half of a days. This clearly indicates instances in which multiple reports were provided in a narrow range within the same sampling interval, i.e. in the evening. Because of this, it is no longer possible to just use the timestamps associated with any report to situate the episodes that they refer to even at a coarse level of half a day.

7.5.2. LIFELOG TIMELINE OF CONTACT DETECTIONS

The dataset contains a range of records that have been obtained through pervasive monitoring of participants' behaviors during their daily social interactions throughout the study. These recordings were collected by devices known as *Sociometric Badges*, wearable monitoring platforms that continuously sense their users' relative motion, acoustic ambiance, and the proximity to other badges. For an in-depth description of all the data captured by such a device we refer the reader to [19].

Of relevance for the current study is that badges create a timeline of annotations that

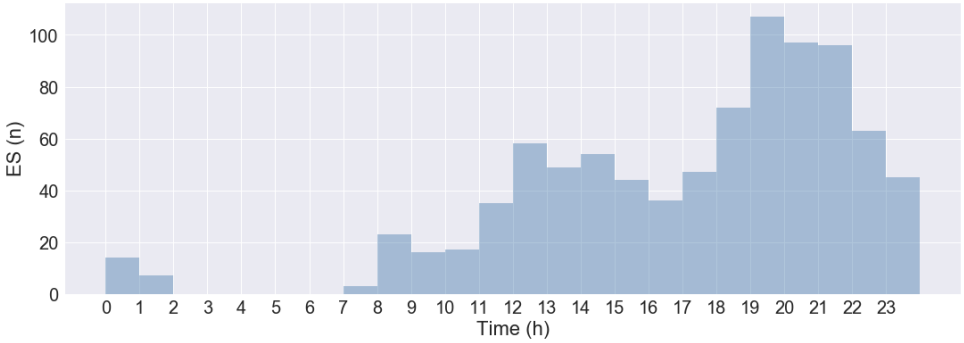


Figure 7.2: Distribution of the time when participants handed in Experience Samples (hours)

Table 7.2: Contact Detections per Participant.

	P1	P2	P3	P4	P6	Total
N	28194	21840	30018	913	13142	94107

uniquely identify any other badge they encounter in close proximity. The devices create this information through a hardware-based process: each device broadcasts a uniquely identifiable infrared signal that can be received by other badges within a reception cone with a 30 degree in a range of about 1.5 meters [20]. Research has demonstrated that this method is reliable at detecting co-location of wearers, but that its ability to do so comprehensively is negatively impacted by barriers and the limited detection range [20]. We will refer to this data for the remainder of this chapter simply as *Contact Detections*.

Table 7.2 provides an overview of the total amount of such contact detections that have been registered by the badges of each participant in the study.

7.6. EMPIRICAL INVESTIGATIONS

In this section we describe a series of empirical investigations in which we model the memory responses of participants in the previously described study to discover representations for the episodes in our dataset.

As an initial step, we identified properties of the episodes that participants have recollected as part of their memory response which a plausible representation should correspond with. For this we turned towards the data that is available as part of their self-reported descriptions, as well as the instructions that they were provided with. We identified the following two attributes:

- **Presence:** there is a part of participants' self-reports that details exactly which other people were present during the episode that they refer to in their experience sample. This means any plausible DERs should involve references to this group of fellow participants.
- **Recency:** when prompted, participants were explicitly instructed to report the most

recent instantiation of what they considered to be a social interaction. Therefor, a plausible representation will need to be situated in temporal proximity to the moment of recall. This moment is documented as part of their self-reported experience samples.

In the following, we first detail how we preprocessed the dataset for usage in our empirical investigations. We then describe the correspondence evaluation function that we modeled for the memory response in this study. Finally, we outline an experiment in which we explore the degree of similarity of the correspondence in the representation that our model is able to discover within- and across-individuals. Finally, we provide an overview of the DERs that our model proposes for the episodes in this dataset.

7.6.1. DATA PREPROCESSING AND SELECTION

In this section we account for how we preprocessed and selected the elements from the dataset that we deemed relevant for discovering representations that display correspondence with participants' recollections in terms of their *presence* and *recency*.

For this purpose, we use information that was provided by participants as part of their experience samples $S = \{s_0, s_1, \ldots, s_n\}$, and the contact detection-data that was recorded by the sociometric badges of participants. We interpret the latter as a lifelog timeline of perception proxies $P = \{p_0, p_1, \ldots, p_m\}$.

Here, a single experience sample is a record s in the form of a 3-tuple (t, R, o). Were, t, refers to an integer timestamp denoting the time at which a sample was handed in by a participant, while R refers to a set of labels that denote which other participants' were reported as being present in the episode referred to by the experience sample. Finally, o is a label denoting the identity of the participant that is the author of the experience sample.

The available data on contact detections form a lifelog timeline of individual records p that are timed perception proxies for the presence of specific other participants. Each instance of such a proxy is also represented in the form a 3-tuple, (t, a, o). The meaning of t is the time at which the record was created, a is the label of the participant that was detected, and o is the label identifying the participant from who's perspective the proxy was recorded.

PREPROCESSING EXPERIENCE SAMPLES

We excluded 22 experience samples from the dataset due to malformed entries, or because they were likely misreports. Additionally, we re-dated some self-reports from within the pool of available samples for participants as part of the preprocessing for our experiment. We modified all reports that were handed in before 3am in the morning and for which no available lifelog data exists for this period from within the same day. In these cases we assumed that a sampling interval had been skipped by participants, i.e. that they had reported an episode from the day before. To more accurately reflect participants' recollection behavior, we associated such samples with the previous day (11:59:59pm).

PREPROCESSING LIFELOG DATA

The perception proxies contained in the lifelog timelines of the study are not mutual. This means that there exist instances where one participant's records indicate contact with another person, without that person's sensor producing a matching entry in their

Table 7.3: Final number of data pairs (P_i^+, s_i) selected for usage in our experiments.

	P1	P2	P3	P4	P6	Total
N	152	146	138	24	112	572

own lifelog timeline. However, we assume that co-location at such close range as is being registered by the wearable sensor result in mutual perceptions between participants (i.e. *"If I see you, then you see me as well"*). To reflect this, we mirrored entries across participants' timelines and combined these mirrored versions with the original lifelogs into an extended dataset P^+. It fulfills the following constraints:

$$\forall p_x \in P^+ \forall p_y \in P^+ \Big(p_x \neq p_y \wedge t(p_x) = t(p_y) \wedge o(p_x) = a(p_y) \wedge a(p_x) = o(p_y) \Big) \quad (7.3)$$

ALIGNMENT AND SELECTION

Finally, we partitioned our lifelog dataset P^+ into individual segments, each spanning a period of time in which representations for a specific experience sample should be discovered. This means from the beginning of the same day on which the episode has occurred, up to the moment it was reported. That is, all parts resulting from this partitioning $\{P_0^+, P_1^+, \ldots, P_n^+\}$ fulfill the following constraints:

$$\forall i \Big(\forall p \in P_i^+ \Big(t(p) < t(s_i) \wedge day(p) = day(s_i) \wedge o(p) = o(s_i) \Big) \Big) \quad (7.4)$$

The result is an aligned dataset that contains parings of participants' experience samples with relevant segments from within their lifelog timelines (P_i^+, s_i).

From the total amount of 861 such pairings, not all did meet our requirements. We removed an additional 220 such pairings, because there was no lifelog data present for the relevant period of time.

Furthermore, we had to remove a set of 70 samples for which there was no overlap between the people that were present in a participant's description of the episode that he/she recollected for the experience sample and the associated lifelog data. Table 7.3 provides an overview of the remaining data pairings that we used for our experiments split by participants.

7.6.2. MODELING PARTICIPANTS' MEMORY RETRIEVAL PROCESSING

In this section we describe our computational approach for assessing the degree with which a lifelog segment displays correspondence with the available information about an episode. As mentioned above, we identified two attributes of the episodes in this scenario that representations will need to meet: the presence of specific other participants, and recency w.r.t. the moment of recall.

To assess the degree to which a lifelog segment corresponds with these properties, we constructed the following evaluation function cor:

$$cor(C,P,s) = pres(C,P,s) * rec(C,P,s) \tag{7.5}$$

where C is a given candidate segment of a participant's lifelog P, s is relevant information about the recollected episode. The total evaluation of a candidate C consists of two partial functions, each of which assesses the degree to which one of the correspondence requirements is met. In our model, plausible representations need to possess both attributes jointly for achieving maximum correspondence.

PRESENCE EVALUATION

For assessing the correspondence between a lifelog segment in terms of the people that were reported as present by a participant, we compared the degree to which the labels of the perception proxies it contains match their description in the following way:

$$pres(C,P,s) = \frac{sim(C,s)}{sim_{max}(P,s)} \tag{7.6}$$

In this function, $sim(C,s)$ is the *Jaccard Similarity* between the set of all annotations describing the presence of participants in the lifelog segment C and the set of labels that denote who was present in the associated self-report s. We normalized this measure over a value generated via the operation $sim_{max}(P,s)$. It denotes the maximum possible overlap between the annotations contained in the lifelog from which the segment under investigation was created$P \supseteq C$ and the self-report s. The reason for this procedure is that there are cases in which not all individuals that were reported as present were also detected within the relevant lifelog. This may be a result of the rather short detection range of the sociometric badges, causing participants not to be registered, even though they are perceived as present. Together, this function provides a relative measure of a segment's correspondence w.r.t. the presence of other participants in the range from $[0,1]$. A 0 denotes a total discrepancy between the two accounts, while a 1 forms the best match possible for a representation created from this particular lifelog timeline.

RECENCY EVALUATION

Next, we devised an evaluation function to assess the degree to which a lifelog segment C under evaluation displays recency w.r.t. the moment at which the memory response took place, as indicated by the timestamp in the associated self-report s:

$$rec(C,P,s) = 1 - \Big((t_{rel}(s,P) - t_{rel}(C,P) \Big) \tag{7.7}$$

In essence, this function provides a measure between the time when a self-report was provided, and the beginning of the lifelog segment under evaluation (i.e. the timestamp of the first perception proxy). Importantly, these moments transformed to their relative position within the timeline of the lifelog from which the segment was created $P \supseteq C$. This is achieved by normalizing both objective timestamps over the duration that is covered by the lifelog timeline, an operation that is denoted by $time_{rel}$. The resulting overall measure for recency for any given segment under evaluation falls within the interval $[0,1]$,

Table 7.4: Average Results for all Experiments

	N	AvgCor$\pm SD$ ($\Delta Train$)	AvgPres$\pm SD$ ($\Delta Train$)	AvgRec $\pm SD$ ($\Delta Train$)
WithinCV	25	$.60 \pm .09$ ($-.02$)	$.73 \pm .07$ ($-.01$)	$.85 \pm .08$ ($-.01$)
StratCV	5	$.59 \pm .03$ ($-.01$)	$.72 \pm .03$ ($-.01$)	$.85 \pm .04$ ($< .00$)
LopoCV	5	$.58 \pm .02$ ($-.02$)	$.73 \pm .07$ ($+.01$)	$.83 \pm .07$ ($-.02$)

where a 0 denotes a maximally distant segment (i.e. it is located at the furthers point away in the timeline of the lifelog), while a 1 is a maximally recent one (it is the closest point in the relative timeline of the lifelog).

7.6.3. EXPLORATION OF SIMILARITY IN REPRESENTATION DISCOVERY

An implicit assumption of our model for the memory responses in this study is that they are highly similar to each other. That is: prompting individuals to remember experiences in their past using the same prompt is assumed to result in a very similar form of recollection for each instance, independently of who is confronted with it, or when that is. Arguably, the existence of such a shared memory response is an essential property for experience sampling. Without it, these methods would not be able to provide comparable information from different participants in a study and at different moments in time.

In this section we explore whether our model would display a behavior that reflects this property when discovering correspondent representations for episodes in this study.

To gain insights into this, we conducted three experiments using our preprocessed dataset in different cross-validation schemes. These allowed us to study the degree to which a model that was trained to reflect the memory responses of some subset of our data, would vary in the correspondence that it produces when being applied to unseen instances.

EXPERIMENTAL SETUP

For the purpose of this exploration we devised the following three cross-validation schemes:

- **WithinCV:** For each participant we partitioned all available pairings of (P_i^+, s_i) into five segments. Each segment was populated via random sampling without replacement. We used this division in a 5-Fold Cross-Validation procedure for training and testing of a model for each individual.

- **StratCV:** This experiment involves training and testing with a 5-Fold Cross-Validation procedure. Each partition is populated by randomly selecting pairings (P_i^+, s_i) without replacement. The amount of pairings that are selected from each participant's data to populate a segment is proportional to their share in the overall amount.

- **LopoCV:** In this experiment, we split all available pairings (P_i^+, s_i) into 5 segments. Each consists of all the data associated with a specific individual in the study. Training is then undertaken in a Leave-One-Participant-Out fashion. That means,

we first train our model on data of 4 participants, and then apply it to the held-out data from the remaining individual.

The goal of the WithinCV-procedure was to gain insights into the similarity of the correspondence-scores produced by our model when trained and tested based on instances belonging to the same individual. In contrast, both StratCV and LopoCV provide insights into the consistency of the correspondence displayed by our model for instances belonging to different participants in the study.

In order to reflect the memory responses underlying the episodes described in the study, we train our model to learn a parameter Δt that maximizes the average correspondence of proposed representations over all available pairs of data (P_i^+, s_i) that were assigned to a particular segment of the training-data:

$$\underset{\Delta t}{\operatorname{argmax}} \frac{1}{n} \sum^{i=n} \Big(\max_{C \in enc(P_i^+, \Delta t)} cor(C, P_i^+, s_i) \Big) \tag{7.8}$$

Since in our scenario the timeline spanned by lifelogs consists of only a single day, we optimized correspondence during training with a sweep Δt over the interval $[0, 20000]$ (seconds). This means, that two consecutive perception proxies in the lifelog cannot not be farther apart than $5\frac{1}{2}$ hours from each other to be counted towards the same segment. In situations where multiple optimal solutions for Δt were discovered in a training phase, we selected the one with the smallest value. This corresponds with a preference for models with a more specific segmentation over broader ones.

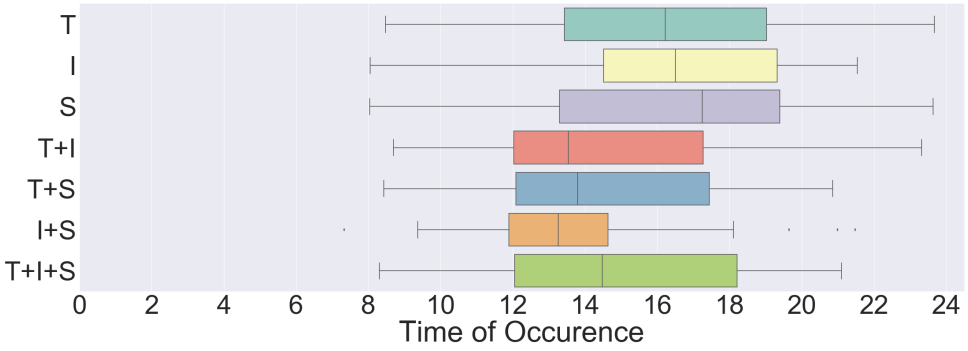
7

RESULTS

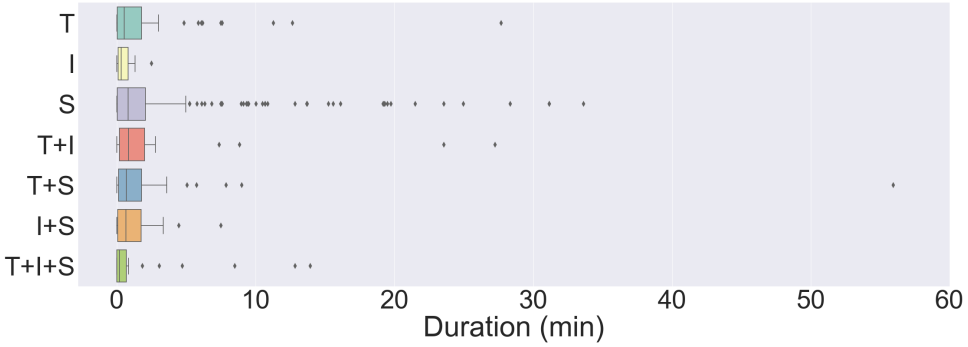
The information in *Table 7.4* represents the average results that were achieved in these experiments (i.e. averaged over all folds). The optimized average correspondence for DERs achieved by our model varied only minimally between the testing and training phases ($\Delta Train$). This is the case independently of whether it was trained to reflect memory responses within a single participant, or when spanning data from different persons. Moreover, both the recency and presence components that comprise these correspondence scores display such a similarity. We see in this behavior a property that one would expect in an experience sampling scenario, i.e. a substantial degree of similarity across all instances of the memory responses. This adds further plausibility to the representations that are discovered by our model for the memory response in this scenario.

7.6.4. EXPLORATION OF DISCOVERED EPISODE REPRESENTATIONS

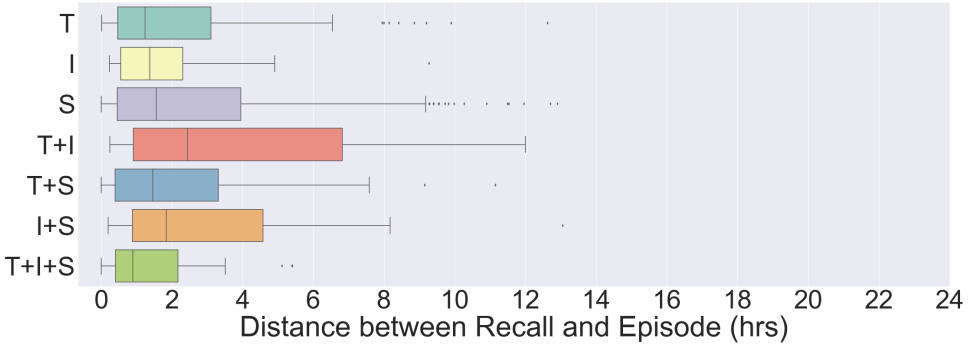
In this section we explore the DERs that were discovered by our model for the recollections of participants in this study when trained in a person-independent fashion on all available pairings (P_i^+, s_i). The discussions in this section are not intended to provide a thorough analysis of participants' social interactions. Instead, they form a demonstration of the insights that possession of DERs could provide to support researchers that undertake such an endeavor.



(a) Distribution of the Time of Occurrences for DERs, sorted by Interaction-Type (in 24h-format).



(b) Distribution of the Duration of discovered DERs(Minutes) by Interaction-Type.



(c) Distribution of the delay between the discovered DERs and the moment of participants' recollection (in Hrs).

Figure 7.3: Discovered Representations by Interaction Type. Labels refer to T: Task Interaction related to Team Goals, I: Task Interaction related to Individual Responsibilities, S: Social Interaction. Labels combined with a '+' represent interactions that were labeled as mixed by participants.

TIME OF OCCURRENCE

Most of the discovered DERs are located in the afternoon (M = 15.753.83, SD = 3.83, N = 571), but they cover the entire waking day period of participants. *Figure 7.3a* describes the distribution of where episode representations are situated, sorted according to how

participants labeled the interactions during them. A potential pattern that can be spotted when looking at this distribution relates to representations for episodes which revolve around a mixture of individual responsibilities and socializing (i.e. I+S-type interactions). These are generally situated at midday ($M = 13.65$, $SD = 3.22$, $N = 30$). This is not the case for representations of episodes that are perceived as either being purely social (S-type interactions, $M = 16.23, SD = 3.86, N = 343$), or to entirely revolve around work (I-type interactions, $M = 16.18$, $SD = 4.01$, $N = 11$). Both of these tend to be situated rather later in the day. Together, this could indicate that the activities spanned by I+S interactions describe meetings where individual tasks were discussed among team-members over shared meals around lunchtime.

DURATION

The average duration of the DERs discovered by our approach was $M = 2.25$ minutes ($SD = 5.12, N = 571$). *Figure 7.3b* describes their distribution according to the associated interaction-type. The discovered DERs for purely social interactions (S-type) tend to have the longest average duration ($M = 2.42$, $SD = 5.05, N = 343$). On the other hand those that were characterized as revolving around individual responsibilities take up the shortest average amount of time ($M = 0.62$, $SD = 0.74, N = 11$). This could be a result of the strongly task-oriented nature that participants ascribe to these interactions, reflecting short and efficient discussions.

DISTANCE TO RECALL

Another interesting aspect of the discovered DERs is their relative distance to the point in time at which participants provided a corresponding self-report (see *Figure 7.3c* for the distribution according to Interaction-Type). On average, representations are situated around three hours before a participant's self-report ($M = 2.71$, $SD = 2.91$, $N = 571$). Such information could, for example, be helpful in identifying an opportune structure for requesting self-reports in a study.

7.7. SUMMARY AND CONCLUSION

The combination of ubiquitous monitoring and self-reported reflections into a synchronized timeline has potential for increasing our understanding of how behavior emerges and unfolds in the scope of everyday lives. We have argued that one principal challenge that needs to be addressed to make progress towards providing such a synchronized description, is to organize data in a fashion that is analogous to how individuals experience their personal past in recollection.

In this chapter, we have suggested that lifelogs form a meaningful source for representations to anchor remembered episodes within an objective timeline. To this end, we have described an approach for discovering candidates for such representations by computationally modeling the memory responses underlying their recollection. We have applied this approach to a dataset describing recollections of participants in a longitudinal study, and have argued that this has resulted in plausible representations for them. Our brief exploration of these representations has hinted at some of the insights that might be gained about individuals' social interactions through their study.

Undertaking our empirical investigation has revealed several opportunities for further exploration. First, while we consider the discovered representations in our scenario as plausible, we did not demonstrate that they are also *accurate*. That is, we have not provided empirical evidence for the degree to which their estimated position in a timeline corresponds with the period referred to by participants when providing a self-report. Future research might explore ways of conducting such evaluations, as well as the collection of relevant data for it. Second, while annotations in a lifelog timeline have the potential to indicate information that *could have* been perceived by monitored individuals, they are not guaranteed to reflect what actually *was* perceived by them. This is primarily caused by their inability to mirror human attentional processes when creating annotations. In our opinion, this forms a general challenge for lifelogging as a research field. A starting point for addressing it may be found in existing research that explores the computational modeling of human attentional processes [21].

In summary, we see our approach as a contribution towards enabling ubiquitous computing applications to create synchronized descriptions that reflect how people experience their daily lives, as well as how they behave in them. In our opinion, the information provided by lifelogs forms a valuable resource for bridging the gap between remembered experience and objectively collected data, and its potential in this respect should be the target of further research.

REFERENCES

[1] C. Napa Scollon, C.-K. Prieto, and E. Diener, *Experience Sampling: Promises and Pitfalls, Strength and Weaknesses,* in *Assessing Well-Being: The Collected Works of Ed Diener*, Vol. 4, edited by E. Diener (Springer, Dordrecht, 2009) pp. 157–180.

[2] J.-P. P. Laurenceau, L. F. Barrett, and P. R. Pietromonaco, *Intimacy as an interpersonal process: the importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges.* Journal of personality and social psychology **74**, 1238 (1998).

[3] S. Shiffman, M. Hufford, M. Hickcox, J. A. Paty, M. Gnys, and J. D. Kassel, *Remember that? A comparison of real-time versus retrospective recall of smoking lapses.* Journal of Consulting and Clinical Psychology **65**, 292 (1997).

[4] S. W. J. Kozlowski, G. T. Chao, C. H. Chang, and R. Fernandez, *Big Data at Work: The Data Science Revolution and Organizational Psychology* (Routledge, 2015) pp. 272–309.

[5] E. Salas, R. Grossman, A. M. Hughes, and C. W. Coultas, *Measuring Team Cohesion: Observations from the Science,* Human Factors: The Journal of the Human Factors and Ergonomics Society **57**, 365 (2015).

[6] E. Tulving, *Episodic Memory: From Mind to Brain,* Annual Review of Psychology **53**, 1 (2002).

[7] M. A. Conway, *Episodic memories,* Neuropsychologia **47**, 2305 (2009).

[8] C. Gurrin, A. F. Smeaton, and A. R. Doherty, *LifeLogging: Personal Big Data,* Foundations and Trends® in Information Retrieval **8**, 1 (2014).

[9] M. L. Lee and A. K. Dey, *Lifelogging memory appliance for people with episodic memory impairment,* in *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp '08*, Vol. 344 (ACM Press, New York, New York, USA, 2008) p. 44.

[10] A. R. Doherty and A. F. Smeaton, *Combining face detection and novelty to identify important events in a visual lifelog,* Proceedings - 8th IEEE International Conference on Computer and Information Technology Workshops, CIT Workshops 2008 , 348 (2008).

[11] C. Gurrin, A. F. Smeaton, Z. Qiu, and A. Doherty, *Exploring the technical challenges of large-scale lifelogging,* in *Proceedings of the 4th International SenseCam & Pervasive Imaging Conference on - SenseCam '13* (ACM Press, New York, New York, USA, 2013) pp. 68–75.

[12] E. Talavera, M. Dimiccoli, M. Bolaños, M. Aghaei, and P. Radeva, *R-Clustering for Egocentric Video Segmentation,* in *Pattern Recognition and Image Analysis*, edited by P. R., C. J., and P. X (Springer International Publishing, 2015) pp. 327–336.

[13] P. Wang and A. F. Smeaton, *Aggregating semantic concepts for event representation in lifelogging,* in *Proceedings of the International Workshop on Semantic Web Information Management - SWIM '11* (ACM Press, New York, New York, USA, 2011) pp. 1–6.

[14] A. R. Doherty, A. F. Smeaton, K. Lee, and D. P. W. Ellis, *Multimodal segmentation of lifelog data,* RIAO 2007 LargeScale Semantic Access to Content Text Image Video and Sound , 21 (2007).

[15] M. A. Conway and C. W. Pleydell-Pearce, *The construction of autobiographical memories in the self-memory system.* Psychological Review **107**, 261 (2000).

[16] S. Farrell, *Temporal clustering and sequencing in short-term memory and episodic memory.* Psychological Review **119**, 223 (2012).

[17] D. Clewett and L. Davachi, *The ebb and flow of experience determines the temporal structure of memory,* Current Opinion in Behavioral Sciences **17**, 186 (2017).

[18] Y. Zhang, J. Olenick, C.-H. Chang, S. W. J. Kozlowski, and H. Hung, *The I in Team,* in *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI '18*, IUI '18 (ACM Press, New York, New York, USA, 2018) pp. 421–426.

[19] T. Choudhury and A. S. Pentland, *The sociometer: A wearable device for understanding human networks,* CSCW'02 Workshop: Ad hoc Communications and Collaboration in Ubiquitous Computing Environments, New Orleans, Louisiana, USA (2002), 10.1.1.57.9810.

[20] D. Chaffin, R. Heidl, J. R. Hollenbeck, M. Howe, A. Yu, C. Voorhees, and R. Calantone, *The Promise and Perils of Wearable Sensors in Organizational Research,* Organizational Research Methods **20**, 3 (2017).

[21] C. Roda and J. Thomas, *Attention aware systems: Theories, applications, and research agenda,* Computers in Human Behavior **22**, 557 (2006).

8

General Discussion

ABSTRACT

In this final chapter, we review this dissertation's goals and the contributions that it makes w.r.t. these. In doing so, we briefly summarize the key findings presented in the relevant chapters and highlight their broader implications. After discussing some general limitations of the presented research, we reflect on open challenges and opportunities for future work.

8.1. CONTRIBUTIONS AND FINDINGS

This dissertation set out to pursue two primary research objectives for progress on recollection-aware predictions of user affect.

G1 the identification of the information that is necessary for a computer system to facilitate recollection-aware modeling of user affect, as well as the additional prediction challenges that need to be solved for providing this information, and

G2 an evaluation of the *effectiveness* and *feasibility* of addressing these prediction challenges based on data available to automated systems in relevant application domains.

In pursuit of these research goals, we have presented the following contributions:

Proposal of a Computational Architecture and Research Framework (Ch. 2): In formulating the RECAP-problem, we identified three primary prediction challenges (predictions of REceptiveness for Recollections, recollected memory Content, and influence of memory content on emotional APraisal) that need to be overcome to achieve recollection-aware modeling of user affect. The proposed computational architecture for an *Artificial Empathic Memory (AEM)* has a double function. First, its components guide research on addressing individual RECAP-challenges with focused technological explorations. At the same time, it offers a psychologically-grounded blueprint that outlines how these individual research efforts link up to contribute to the larger goal of recollection-awareness in predicting user affect.

A Multimodal Dataset for Modeling Affect and Memory Processes (Ch. 3): For studying and modeling the impact of recollected personal memory content on emotional appraisal, we have collected a rich multimodal dataset using crowd-sourcing. It captures individuals' emotional responses to video content, a highly relevant application domain for automatic predictions of user affect. The corpus is the first resource that facilitates research on memory-affect interactions. It can be of great value for future use in the affective computing community (see Ch. 3 for details).

Exploration of the Appraisal-Challenge in Automatic Predictions of Video-induced Emotions (Ch. 4, Ch. 5 and Ch. 6): While predicting responses to video content is of great interest for Affective Computing research, direct empirical insights into how the recollection of personal memory content relates to their emotional impact were not available in prior research. Through a series of statistical analyses of the Mementos dataset (Ch.

4), we have demonstrated differences in how videos are experienced when recollections occur and identified a strong connection to the recollected memory content's emotional interpretation. These findings indicate the relevance of personal memories as a contextual driver for video experiences and provide novel insights for developing affective computing applications and media psychology in general. In machine learning experiments, we show that automatic analysis of self-reported memory descriptions is a feasible approach for computational modeling their impact on individuals' emotional appraisal of video stimuli (Ch. 5). Furthermore, we investigate personal memories' effectiveness as context for explaining variation in emotional responses to stimuli alongside existing strategies (user profiles, audiovisual content analysis, and facial behavior analysis). Personal memories substantially improve models that rely on viewers' static attributes, such as demographics, personality, and mood for predictions (Ch. 4). Moreover, automatic analysis of memory descriptions complements and improves predictions based on the analysis of videos' audiovisual content (Ch. 5) or viewers' facial expressions (Ch. 6). Importantly, our findings indicate that context-free analysis of either video content (Ch. 5) or facial expression (Ch. 6) offers only very limited insights into variation in individuals' subjective emotional experiences. Beyond highlighting the importance of recollection-aware predictions, they also underline the need to explore context-sensitive approaches to affect prediction more generally.

Exploration of the Content Challenge based on Lifelog Data (Ch. 7): A crucial part for addressing the challenge of content is the availability of information for computational models about individuals' past that might be recollected as memories (i.e., an episodic store in the AEM architecture, see Ch.3). Lifelogging technology is a form of ubiquitous data collection that holds the potential for building such a comprehensive repository about a person's daily activities, creating a timeline of semantic annotations documenting aspects of a person's impressions in sequence (e.g., people, places, objects). However, a particular challenge for using such a lifelog for an AEM is that the estimated content of memories should match what is remembered by an individual in a particular recollection instance. Such a match is difficult to establish because of a challenge that we have dubbed the *recollection gap* (Ch. 7). Data in a lifelog is typically organized as sequential entries in an objective timeline. In contrast, remembered episodes are not experienced as having static temporal boundaries but are dynamically reconstructed at recall in response to cues and referenced by their narrative content and context. However, the semantic annotations in a lifelog may hold the potential to bridge this gap and provide a plausible alignment to the experiential content of a memory. We explore this by modeling individuals' recollection processes when cued with a particular prompt about their most recent social interactions based on existing lifelog data from a longitudinal social science study (Ch. 7). Our empirical investigations show the plausible behavior of our models and their episode retrieval for the scenario captured by the dataset. However, while plausible, these findings are tentative and cannot be validated against participants' actual experiences with the available dataset. Thus, while the study hints at the potential of extracting meaningful representations for remembered events from lifelogs, it points to the need for specifically created datasets for further computational research and challenges for its validation.

8.2. PRACTICAL IMPLICATIONS

Free-Text Memory Descriptions are a valuable resource for Affect Prediction: A direct practical implication of our findings for modeling affective responses to media is that free-text memory descriptions may offer a rich resource for predictive modeling. As such, purposeful mining from available user data (e.g., social media comments) may warrant further investigation and can likely benefit prediction applications directly. Moreover, collecting descriptions of personal memories through triggers for stimulating reminiscence experiences may not only be an engaging activity for users in itself [1] but also offer data for context-sensitive predictions of emotional responses later on.

Assumptions of Video Affective Content Analysis need to be considered: Video Affective Content Analysis (VACA) is the endeavor to build systems that can estimate the emotional impact that a video stimulus will have on viewers [2]. This task is traditionally approached by analyzing only the video's audiovisual content in isolation and assigning a single label to viewers' expected affective response[3]. Our findings provide quantitative evidence highlighting the limitations of modeling affective responses to media content as homogeneous and context-free in this way (see Ch. 3-5). There can be substantial variation in responses to any specific stimulus (within and across individuals). Consequently, practical applications making these assumptions should consider these limitations when relying on this strategy. Ideally, all predictions of media-induced affect would account for personal memories and other relevant contextual influences. However, it is not always possible to provide such specific information to automatic systems in the real world. Importantly, our analysis of the Mementos data (see Ch. 3) suggests that context-sensitivity may not always be necessary and that for some video stimuli, emotional responses do, indeed, display a strong coherence. For such stimuli, traditional VACA assumptions are more tenable than for ones with an overall high degree of variation. Consequently, our findings imply that for purely content-centric analysis to be valid, automatic systems should estimate the degree of variation in affect caused by a specific video. Such an attempt could take the form of using content features to explicitly model the distribution of affective responses to a stimulus (see the work of He&Jin [4] for an example applied to the image domain) as a basis for deciding whether a single expected emotion forms a viable approximation.

Strategic selection of information sources for automatic affect prediction is possible: Analysis of visual data describing facial behavior is perhaps the most widely used approach to automatic affect prediction. However, there are numerous situations where people may not express their feelings or in which collecting such data is not an option (e.g., occlusions or user concerns). Apart from improved accuracy, this is a driving force for developing systems that facilitate multimodal affect predictions. Our results on contextualizing the analysis of facial behavior with that of video content and free-text memory descriptions (see Ch. 6) demonstrate the benefits of this approach. Importantly, however, these different information sources also appear to provide insights about particular aspects of affective experience. While the descriptions of personal memories viewers associated with a video they are watching facilitates better predictions of the pleasure they experience, its content facilitates better predictions of their experienced

dominance. Facial behavior, in contrast, facilitated the best predictions of arousal in our experiments. Together, this suggests that developers can specifically select a mix of information sources for their affect prediction system that is most relevant for achieving accuracy on a particular aspect of affective experience.

8.3. OVERALL LIMITATIONS

We now turn towards a discussion of the overall limitations of the research presented in this dissertation. While each of the chapters presenting empirical work addresses the limitations of its findings, we focus here more generally on the limitations of the presented work for addressing the RECAP problem with an Artificial Empathic Memory.

First, the dataset that we have collected to explore the appraisal-challenge (Ch. 3) captures just a single example from within the particular domain of human-media interactions. There are good reasons to argue that even for responses to video stimuli, the importance of personal memories may differ depending on the content format involved (see Ch. 4 for a discussion). These are valid points and limit our contributions to generalize to a broader area of human affective responses. However, our studies about the effectiveness of accounting for personal memories in predictions of video-induced emotions (Ch. 4-6) do provide valid examples for (1) the relevance of recollection-awareness and (2) the fundamental limitations of existing technical approaches that largely neglect or suppress such contextual influences. Our studies provide clear evidence that justifies further technological research on context-sensitive affect detection more generally and the components of an Artificial Empathic Memory specifically. Naturally, this will require explorations of memory-influences on user affect in other domains.

Secondly, our attempts at computational modeling the appraisal of memory content (Ch. 5+6) have relied on explicit descriptions written down by participants after exposure to the triggering stimulus. Requiring such descriptions limits our findings' immediate relevance for applications since they may not be typically available to many technological systems at prediction time. However, a similar case can be made about facial expressions or physiological signals. Moreover, it is plausible to assume that mining similar personal memory descriptions is possible from comments on social media videos. Nevertheless, unlocking the full potential of analyzing personal memory content for affect modeling in applications requires technology to address the challenges of content and receptiveness with an Artificial Empathic Memory in future research.

8.4. OPEN CHALLENGES AND FUTURE RESEARCH

Modeling of Appraisal Processes Involving Personal Memories: The findings from our empirical investigations support a close connection between the outcome of individuals' overall appraisal of a stimulus event (i.e., how they feel about a video that has been watched) and an appraisal outcome of associated memory content (i.e., how people feel about what a video has reminded them of). However, our findings shed no light on the underlying mechanisms of either how memory content is appraised or how it influences the overall appraisal of a stimulus or event. In general, recollected personal memories (especially past events) seem to contain multiple aspects that individuals can appraise (see Chapter 2 for a discussion), requiring a careful and systematic approach

to modeling them. Concretely, when individuals experience events, appraisal theories posit that these are continuously appraised at different levels of cognitive automaticity [5]. The outcome of this initial appraisal is likely stored as part of the memories for later recollection. Additionally, however, people appraise remembered events from their past in a dynamic fashion when remembering them and can change their appraisals over time (see, e.g., the findings from Levine et al. [6]). Finally, recollected memories relate to the appraisal of ongoing stimulus events (e.g., as indicated by our findings in Ch. 4). To the best of our knowledge, cognitive appraisal theories currently offer no detailed picture of how these different processes relate to each other. Interdisciplinary research efforts should be targeted at untangling these processes and provide a clear theoretical foundation because this can provide payoffs in the form of a deeper understanding of human psychology and behavior, as well as technological progress. Computational modeling of emotions in Affective Computing research (e.g., similar to architectures like EMA [7], or MAMID [8]) can offer an important contribution to that end because it enforces conceptual clarity of relationships and processes and thus supports systematic theory-building [9, 10]. These insights can then provide empirical research on human psychology with a clear hypothesis space for investigations. In turn, a more refined understanding of how memories are appraised, how stable this appraisal is over time and circumstances, and how recollected content influences appraisal of ongoing situations is of direct importance for recollection-aware predictions of user affect. For example, we have discovered video-specific variations in the degree to which the outcome of memory appraisal corresponds to that of video appraisal (see Ch. 4). Thus, estimating the impact of memories on how a stimulus is appraised could help applications better quantify the degree to which recollection-awareness may be relevant for predictions in the first place.

Collecting data for a systematic exploration of memories' role in appraisal processes is likely to pose a substantial challenge. The primary reason for this is the holistic and interdependent nature of human cognition and its relation to self-report. It is difficult for people to reflect on the precise reasons for their emotional responses, and their ability for accurate causal attributions are limited (as demonstrated clearly in a seminal study by Schachter&Singer [11]). Moreover, asking them to break down their overall experience of a situation with a series of nuanced affective evaluations about how different sources contribute to their overall situation can be subject to priming effects. Such a breakdown may be particularly challenging for personal memories, where findings both indicate that they can cause emotional responses [12], as well as be primed by affective states [13]. As such, carefully crafted experimental protocols in laboratory settings may be called for in initial explorations before collecting large-scale corpora for modeling with higher ecological validity.

Multimodal Modeling of Memory Receptiveness and Content: The bulk of this dissertation addresses the appraisal-challenge of the RECAP problem in predictions of affective responses to video content. However, understanding when individuals are receptive to involuntary memories (receptiveness-challenge) and why the content of these memories is what it is (content-challenge) are crucial for practical applications. Moreover, our findings highlight that even partial solutions to these challenges can already contribute to better affective modeling in their own right in the video-viewing scenario. For example,

we identified differences in videos' probability to evoke personal memories (see Ch.2). Distinguishing whether a given video is more or less likely to result in personal memories may help approaches that primarily rely on a context-free analysis of their audiovisual content to provide more accurate predictions of affect (see Ch. 4 for a more extensive discussion).

A crucial challenge for further research on this topic is that no specialized corpora exist that capture these processes for computational modeling [1]. The constrained setting of media responses is fruitful for creating such datasets for initial explorations of modeling memory receptiveness and content. It limits the degree of potential additional contextual influences on memory and affective processing and simplifies data collection procedures (e.g., by lending itself to online surveys). Consequently, initial efforts to construct corpora for these topics could build on the protocol used to construct Mementos (see Ch. 3). For example, it could be altered to better address the content-challenge by involving many repeated measures from individual participants to more stimuli. This modification would allow a better exploration of the relationship between types of triggers and what is being remembered. Similarly, the Mementos protocol could be modified to capture stimulus-specific differences in the capacity to trigger memories (i.e., their evocativeness) by collecting responses to a greater variety of triggers. However, a better approach for gathering data on the conditions for recollections to occur would likely follow protocols similar to those used in mind-wandering research (e.g., close to the protocol used by Pelagatti et al. [14]). Here, individuals provide self-reports immediately when they become aware that their attentional focus shifts away from some external stimulus towards internal thoughts. Such annotation schemes would allow for data capturing a closer coupling between triggering cues and memory evocation.

Finally, a promising endeavor for future research is to address the content-challenge with lifelog data as a resource for recollected memory content. A significant challenge becoming evident from our explorations on this topic (see Ch. 7) is the dynamic and flexible shape of memory content inherently tied to specific recollection acts and the context they occur in [15]. While this is an area of active research in psychology, evidence indicates that a person's history is not composed of a static array of discrete events retrieved like documents. Instead, memories are dynamically assembled from knowledge in a hierarchy of nested (and potentially overlapping) organizational units, such as short episodes and extended life periods [16]. Moreover, even this organization may manifest itself differently between voluntary and involuntary recollections. As such, approaches that are prominent for annotating and organizing lifelog data in technological research – e.g., segmenting a timeline of collected data into fixed, atomic events – are unlikely to be viable for turning them into a resource for modeling memory content in recollection (see also related discussions of the need for such a dynamic organization from a technological perspective by Gurrin et al. [17, 18]). Similarly, the annotations and segmentations collected by visual inspection of interpretable lifelog data (e.g., ego-centric photos)) may better approximate perceptual processes at encoding than organization at the recollection. Corpora will need to account for memory content as a dynamic process and describe what was remembered and under which circumstances to provide data useful for modeling recol-

[1] The Mementos dataset can be useful for some future research on the evocativeness of stimuli, however. See Ch. 3 for a detailed discussion.

lections. As is demonstrated by our approach for addressing the content-challenge with lifelog-data (Ch. 7), the absence of ground truth ratings associating data (e.g., in time) with retrieved memory content as experienced upon recollection makes validation of models difficult. While challenging, the development of datasets that capture this relationship in a psychologically valid and pliable way for machine learning is an important target for future research modeling human memory recollection. One possible way to approach this creation could be within a Learning to Rank-framework for information retrieval [19]. For example, corpora may be created by first asking lifelogging users to describe the memories triggered in them by a prompt (and self-report on the context in which these were triggered). Then they could be presented with interpretable data describing candidates of predicted memories (e.g., using images from a wearable camera). They can then rank these options according to how well they match what they actually remembered and provide further annotations only for the correctly indicated memory content, such as their duration on an objective timeline to improve future selections. This approach to online data collection could form a tractable approach for creating corpora of recollections in the experience sampling framework that we explored in our work (See Ch. 7).

Exploration of Context Variables in Automatic Affect Prediction: We have shown overall benefits of integrating contextual information for automatic predictions of emotional responses to videos (see Ch. 5): they become overall more accurate and robust. This finding points towards significant room for improvement in modeling user affect by accounting for contextual influences, even in the already relatively constrained video-viewing setting captured by the Mementos corpus (see Ch. 3). Given that present-day mobile devices facilitate media consumption everywhere, real-life responses are subject to a much greater degree of variation in their circumstances. Consequently, the benefits of detecting differences in context and account for its effects on responses are likely even greater than suggested by our results. Importantly, while we have covered some variables considered to capture relevant contextual influences for predicting affect – Video Content, Facial Behavior, Demographics, Personality, Mood, and Personal Memories (see Ch. 4-6) –, many more should be investigated for their benefits. Promising targets for initial explorations are user-characteristics that might feasibly discriminate between different individuals' lived experience, such as their gender, their occupation, or their cultural embedding (i.e. ethnicity, or membership in subcultures). These are already implicitly acknowledged as relevant within the affective computing research community [20], remain relatively stable over time, and can likely be applied across different domains and applications. Information capturing individuals' enduring personal values more directly could be especially fruitful since they provide some cross-situational estimate of their motivations relating to emotional appraisals [21]. Despite the potential for robustness and accuracy offered by integrating such person-specific information, a priority for real-world applications is doing so ethically, considering issues of privacy and user control, as well as fairness and inclusiveness.

We have argued at the outset of this dissertation that the widespread uptake of context-sensitive affect prediction faces two main hurdles: (1) a clear conceptual understanding of the structure of context that is relevant for improving the effectiveness of automatic

affect predictions and (2) feasible approaches for providing technological systems with an awareness of it. Future research can benefit from a close collaboration with the social sciences to systematically identify and assess potentially relevant contextual influences on human affective processes. A parallel research line can then strive to identify technological solutions to providing relevant information about these contextual influences to systems for predictions (e.g., by mining social media content to infer an individuals' personality [22]). A core component of this collaborative endeavor is creating valid datasets that also facilitate computational modeling (see also our arguments in Ch. 3). The research methodology that we have followed in this dissertation to address the appraisal-challenge as part of accounting for the contextual influence of personal memories can be seen as an instance of this endeavor.

Responsible Development and Supporting Ethical Use: Besides the questions of effectiveness and technological feasibility explored in this dissertation, recollection-aware modeling of user affect brings with it a range of potential threats arising from how it is *implemented* and, in particular, how it is *used* by applications. One crucial threat results from the intimate nature of an AEM's episodic memory store (see also our discussion on the Content challenge in Ch 1). Consequently, trust in what is collected, how it is stored, and whom it is shared with is an essential issue. Similarly, when deployed in applications, they may use a working AEM to adapt or constrain their functionality in subtle and non-obvious ways, e.g., to provide product recommendations based on the expected impact of associated memories. Without transparency, faulty behavior might constitute a mysterious source of missed opportunities or even harm for a user, e.g., in situations where the episodic contents of the model are in substantial misalignment from their actual recollections. Finally, given the potency of memories as a driver of human emotion and behavior, recollection-aware affect prediction holds tremendous potential for misuse in applications. For example, the technology might be abused by creating media interventions that evoke personal fears based on past experiences to nudge people towards particular economic (buy specific products) or political behaviors (vote for a particular party). Such misuse is a credible possibility, given that large corporations and governmental agencies are precisely those institutions that are presently most likely to possess the resources to construct some version of a working AEM feasibly. Together, addressing these threats alongside the technological hurdles will need to be a primary target of future research efforts. Ideally, these will involve efforts to shape policies and develop community guidelines to steer responsible development. In addition, technological research needs to focus on functionalities that empower users, ensuring that they can inspect, interpret, and curate their AEM in meaningful ways as an additional safeguard against malfunctions and abuse. Such efforts could, for example, draw on existing ideas for managing long-term user models (e.g., through mechanisms of user-controlled forgetting [23]) and personal data stores (e.g., through facilitating encapsulation and user-defined licensing of their data [24]).

REFERENCES

[1] D. Cosley, V. Schwanda, S. T. Peesapaty, J. Schultz, and J. Baxter, *Experiences with a publicly deployed tool for reminiscing,* in *Proc. First Int'l Workshop on Reminiscence Systems* (2009) pp. 31–36.

[2] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, *Affective Video Content Analysis: A Multidisciplinary Insight,* IEEE Transactions on Affective Computing **9**, 396 (2018).

[3] S. Wang and Q. Ji, *Video Affective Content Analysis: A Survey of State-of-the-Art Methods,* IEEE Transactions on Affective Computing **6**, 410 (2015).

[4] T. He and X. Jin, *Image Emotion Distribution Learning with Graph Convolutional Networks,* in *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (ACM, New York, NY, USA, 2019) pp. 382–390.

[5] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, *Appraisal Theories of Emotion: State of the Art and Future Development,* Emotion Review **5**, 119 (2013).

[6] L. J. Levine, V. Prohaska, S. L. Burgess, J. A. Rice, and T. M. Laulhere, *Remembering past emotions: The role of current appraisals,* Cognition & Emotion **15**, 393 (2001).

[7] S. C. Marsella and J. Gratch, *EMA: A process model of appraisal dynamics,* Cognitive Systems Research **10**, 70 (2009).

[8] E. Hudlicka, *Modeling the Mechanisms of Emotion Effects on Cognition.* in *AAAI Fall Symposium: Biologically inspired cognitive architectures* (2008) pp. 82–86.

[9] R. Reisenzein, E. Hudlicka, M. Dastani, J. Gratch, K. Hindriks, E. Lorini, and J.-J. C. Meyer, *Computational Modeling of Emotion: Toward Improving the Inter- and Intradisciplinary Exchange,* IEEE Transactions on Affective Computing **4**, 246 (2013).

[10] J. Broekens, T. Bosse, and S. C. Marsella, *Challenges in computational modeling of affective processes,* IEEE Transactions on Affective Computing **4**, 242 (2013).

[11] S. Schachter and J. Singer, *Cognitive, social, and physiological determinants of emotional state.* Psychological Review **69**, 379 (1962).

[12] C. Mills and S. D'Mello, *On the Validity of the Autobiographical Emotional Memory Task for Emotion Induction,* PLoS ONE **9**, e95837 (2014).

[13] G. E. Matt, C. Vázquez, and W. K. Campbell, *Mood-congruent recall of affectively toned stimuli: A meta-analytic review,* Clinical Psychology Review **12**, 227 (1992).

[14] C. Pelagatti, P. Binda, and M. Vannucci, *A closer look at the timecourse of mind wandering: Pupillary responses and behaviour,* PLoS ONE **15**, 1 (2020).

[15] S. Bluck, N. Alea, and B. Demiray, *You Get What You Need,* in *The Act of Remembering* (Wiley-Blackwell, Oxford, UK, 2010) pp. 284–307.

[16] M. A. Conway and C. W. Pleydell-Pearce, *The construction of autobiographical memories in the self-memory system.* Psychological Review **107**, 261 (2000).

[17] C. Gurrin, A. F. Smeaton, Z. Qiu, and A. Doherty, *Exploring the technical challenges of large-scale lifelogging,* in *Proceedings of the 4th International SenseCam & Pervasive Imaging Conference on - SenseCam '13* (ACM Press, New York, New York, USA, 2013) pp. 68–75.

[18] C. Gurrin, A. F. Smeaton, and A. R. Doherty, *LifeLogging: Personal Big Data,* Foundations and Trends® in Information Retrieval **8**, 1 (2014).

[19] T.-Y. Liu, *Learning to Rank for Information Retrieval* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).

[20] B. Dudzik, M.-P. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong, *Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases,* in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE, 2019) pp. 206–212.

[21] R. M. A. Nelissen, A. J. M. Dijker, and N. K. de Vries, *Emotions and goals: Assessing relations between values and emotions,* Cognition & Emotion **21**, 902 (2007).

[22] J. Golbeck, C. Robles, and K. Turner, *Predicting personality with social media,* in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (ACM Press, New York, New York, USA, 2011) p. 253.

[23] D. Barua, J. Kay, B. Kummerfeld, and C. Paris, *Theoretical foundations for user-controlled forgetting in scrutable long term user models,* Proceedings of the 23rd Australian Computer-Human Interaction Conference on - OzCHI '11 , 40 (2011).

[24] R. Mortier, T. Lodge, T. Brown, D. McAuley, C. Greenhalgh, J. Zhao, J. Crowcroft, L. Wang, Q. Li, H. Haddadi, Y. Amar, A. Crabtree, and J. Colley, *Personal Data Management with the Databox,* in *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking - CAN '16* (ACM Press, New York, New York, USA, 2016) pp. 49–54.

SUMMARY

Automatic affect prediction holds the potential to facilitate computer systems that can display intelligent and adaptive behaviour in application domains that require understanding human thoughts and feelings. However, a substantial challenge for successful affect prediction is the inherently subjective nature of human emotions: the high degree of variation in how they are elicited, experienced, and expressed across different contexts make accurate estimations difficult. While research from the social sciences have identified a broad array of contextual variables that drive such variations, accounting for these with context-sensitive approaches to automatic predictions has only been tentatively explored.

An important and omnipresent contextual influence on our emotional and cognitive interpretation of current events, including interactions with technology, are our personal memories. However, current approaches for automatically predicting users' affective states and responses neither consider whether memories are triggered nor the emotional interpretations of those memories. In this dissertation, we argue that this *recollection-unawareness* is a serious limitation because it potentially prevents computer systems from correctly estimating and interpreting users' affect. Motivated by the potential benefits for personalization of a broad range of human-computer interactions, the primary objectives of this body of work are:

G1 the identification of the information that is necessary for a computer system to facilitate recollection-aware modelling of user affect, as well as the additional prediction challenges that need to be solved for providing this information, and

G2 the evaluation of the *effectiveness* and *feasibility* of addressing these prediction challenges in particular application domains.

To address the first goal, we develop a psychologically-inspired architecture for an Artificial Empathic Memory (AEM) that enables computer systems to simulate personal memories' impact on individual users' affective experiences. It provides a conceptual decomposition of the overall problem into a series of individual functional components. Each of these components solves a specific prediction challenge contributing to the overall task of recollection-aware affect prediction. We summarily refer to these challenges as the *RECAP* problem, comprising of predicting *when* memories are likely to be triggered in (challenge of detecting *RE*ceptiveness), *what* a memories content is (challenge of predicting *C*ontent), and *how* recollected memory content impacts emotional experience (challenge of predicting *AP*praisal). Importantly, this framework facilitates targeted technological research on addressing personal memories' contextual influence in automatic affect predictions.

Building on this, the remainder of the dissertation revolves around the second goal, exploring benefits and approaches for addressing the RECAP problem's aspects. Concretely,

our investigations focus on identifying the *effectiveness* and *feasibility* of incorporating information about individuals' appraisal of recollected memory content into systems for automatic prediction of viewers' responses to video content. For this purpose, we collect a large-scale dataset capturing instances in which videos trigger personal memories in viewers and their affective experience. It includes detailed self-reports about viewers' background, their memories and affective experiences, as well as audiovisual recordings of their facial behaviour. Analysis of the collected self-reports reveals that differences in how viewers' feel about the triggered memories explains a substantial degree of variation in their responses to videos that triggered those memories. Moreover, in a series of machine learning experiments, we show that automatic analysis of free-text memory descriptions is a feasible approach for computational modelling their impact on individuals' emotional appraisal of video stimuli. Following up on this, we investigate the effectiveness of analyzing such descriptions as context for explaining variation in emotional responses to stimuli alongside existing strategies (user profiles, audiovisual content analysis, and facial behaviour analysis). Personal memories substantially improve models that rely on viewers' static attributes, such as demographics, personality, and mood for predictions. Similarly, automatic analysis of memory descriptions complements and improves predictions based on the analysis of videos' audiovisual content or viewers' facial expressions. Beyond highlighting the importance of recollection-aware predictions, our findings in this setting underline the need to explore context-sensitive approaches to affect prediction more generally.

The final study in the thesis presents a tentative exploration of addressing the content challenge of the RECAP problem. A crucial part of this challenge is the availability of information for computational models about individuals' past that might be recollected as memories. Lifelogging technology is a form of ubiquitous data collection that holds the potential for building a comprehensive repository about a person's daily activities, creating a timeline of semantic annotations documenting aspects of a person's impressions in sequence (e.g., people, places, objects). In an empirical study, we explore identifying segments in an objective timeline that plausibly correspond to the subjectively experienced memory content reported by participants in existing data from a longitudinal social science study. While our analyses indicate that our model behaves plausibly for the investigated setting, the dataset does not contain ground truth annotations that facilitate empirical validation of their accuracy. Importantly, investigations point towards the need for specialized datasets that facilitate such modelling and provide insights into the challenges that need to be overcome for creating these.

Overall, the findings of the dissertation point towards the importance of considering personal memories in automatic affect predictions and highlight the potential benefits of doing so in existing technological systems. Moreover, the developed Artificial Empathic Memory-architecture provides a blueprint that can structure and guide future research on the subject. More generally, our studies combining personal memories with other contextual variables offer empirical evidence for the overall value of context-sensitive approaches to improve the accuracy and robustness of automatic affect prediction.

SAMENVATTING

Automatische affectvoorspelling biedt het potentieel om computersystemen te faciliteren die intelligent en adaptief gedrag kunnen vertonen in application domeinen die het begrijpen van menselijke gedachten en gevoelens vereisen. Een substantiële uitdaging voor succesvolle affectvoorspelling is echter de inherent subjectieve aard van menselijke emoties: de hoge mate van variatie in hoe ze worden opgewekt, ervaren en uitgedrukt in verschillende contexten, maakt nauwkeurige schattingen moeilijk. Hoewel onderzoek uit de sociale wetenschappen een breed scala aan contextuele variabelen heeft geïdentificeerd die dergelijke variaties veroorzaken, is het slechts voorlopig onderzocht om hiermee rekening te houden met contextgevoelige benaderingen van automatische voorspellingen.

Een belangrijke en alomtegenwoordige contextuele invloed op onze emotionele en cognitieve interpretatie van actuele gebeurtenissen, inclusief interacties met technologie, zijn onze persoonlijke herinneringen. De huidige benaderingen voor het automatisch voorspellen van de affectieve toestanden en reacties van gebruikers houden echter geen rekening met de vraag of herinneringen worden geactiveerd, noch met de emotionele interpretaties van die herinneringen. In deze dissertatie stellen we dat deze *recollection-unawareness* een serieuze beperking is, omdat het computersystemen mogelijk verhindert om het affect van gebruikers correct in te schatten en te interpreteren. Gemotiveerd door de potentiële voordelen voor personalisatie van een breed scala aan mens-computer-interacties, zijn de belangrijkste doelstellingen van dit werk:

G1 de identificatie van de informatie die nodig is voor een computersysteem om herinneringsbewuste modellering van gebruikerseffect te vergemakkelijken, evenals de aanvullende voorspellingsuitdagingen die moeten worden opgelost om deze informatie te verstrekken, en

G2 de evaluatie van de *effectiviteit* en *haalbaarheid* van het aanpakken van deze voorspellingsuitdagingen in bepaalde application domeinen.

Om het eerste doel te bereiken, ontwikkelen we een psychologisch geïnspireerde architectuur voor een Artificial Empathic Memory (AEM) waarmee computersystemen de impact van persoonlijke herinneringen op de affectieve ervaringen van individuele gebruikers kunnen simuleren. Het biedt een conceptuele decompositie van het totale probleem in een reeks individuele functionele componenten. Elk van deze componenten lost een specifieke voorspellingsuitdaging op die bijdraagt aan de algemene taak van herinneringsbewuste affectvoorspelling. We noemen deze uitdagingen samenvattend het *RECAP*-probleem, bestaande uit het voorspellen van *wanneer* herinneringen zullen worden geactiveerd in (uitdaging om *RE*ceptiveness te detecteren), *wat* een herinneringsinhoud is (uitdaging bij het voorspellen van *C*ontent), en *hoe* herinnerde geheugeninhoud de emotionele ervaring beïnvloedt (uitdaging bij het voorspellen van *AP*praisal). Belangrijk

is dat dit raamwerk gericht technologisch onderzoek mogelijk maakt om de contextuele invloed van persoonlijke herinneringen bij automatische affectvoorspellingen aan te pakken.

Hierop voortbouwend, draait de rest van het proefschrift om het tweede doel, het verkennen van voordelen en benaderingen voor het aanpakken van de aspecten van het RECAP-probleem. Concreet richten onze onderzoeken zich op het identificeren van de *effectiviteit* en *haalbaarheid* van het opnemen van informatie over de beoordeling door individuen van herinnerde geheugeninhoud in systemen voor automatische voorspelling van de reacties van kijkers op video-inhoud. Voor dit doel verzamelen we een grootschalige dataset die gevallen vastlegt waarin video's persoonlijke herinneringen oproepen bij kijkers en hun affectieve ervaring. Het bevat gedetailleerde zelfrapportages over de achtergrond van kijkers, hun herinneringen en affectieve ervaringen, evenals audiovisuele opnamen van hun gezichtsgedrag. Analyse van de verzamelde zelfrapportages laat zien dat verschillen in hoe de kijkers denken over de getriggerde herinneringen een aanzienlijke mate van variatie verklaren in hun reacties op video's die die herinneringen hebben geactiveerd. Bovendien laten we in een reeks machine learning-experimenten zien dat automatische analyse van vrije-tekstgeheugenbeschrijvingen een haalbare benadering is voor het computationeel modelleren van hun impact op de emotionele beoordeling van video-stimuli door individuen. In vervolg hierop onderzoeken we de effectiviteit van het analyseren van dergelijke beschrijvingen als context voor het verklaren van variatie in emotionele reacties op stimuli naast bestaande strategieën (gebruikersprofielen, audiovisuele inhoudsanalyse en analyse van gezichtsgedrag). Persoonlijke herinneringen verbeteren aanzienlijk modellen die afhankelijk zijn van de statische kenmerken van kijkers, zoals demografie, persoonlijkheid en stemming voor voorspellingen. Evenzo vult automatische analyse van geheugenbeschrijvingen de voorspellingen aan en verbetert deze op basis van de analyse van de audiovisuele inhoud van video's of de gezichtsuitdrukkingen van kijkers. Naast het benadrukken van het belang van recollection-aware voorspellingen, onderstrepen onze bevindingen in deze setting de noodzaak om contextgevoelige benaderingen te onderzoeken om voorspelling meer in het algemeen te beïnvloeden.

De laatste studie in het proefschrift presenteert een tentatieve verkenning van het aanpakken van de inhoudelijke uitdaging van het RECAP-probleem. Een cruciaal onderdeel van deze uitdaging is de beschikbaarheid van informatie voor computermodellen over het verleden van individuen die als herinneringen kunnen worden herinnerd. Lifelogging-technologie is een vorm van alomtegenwoordige gegevensverzameling die het potentieel biedt voor het bouwen van een uitgebreide opslagplaats over de dagelijkse activiteiten van een persoon, waarbij een tijdlijn van semantische annotaties wordt gecreëerd die aspecten van iemands indrukken in volgorde documenteren (bijv. mensen, plaatsen, objecten). In een empirische studie onderzoeken we identificerende segmenten in een objectieve tijdlijn die aannemelijk overeenkomen met de subjectief ervaren geheugeninhoud die is gerapporteerd door deelnemers aan bestaande gegevens van een longitudinaal sociaalwetenschappelijk onderzoek. Hoewel onze analyses aangeven dat ons model zich plausibel gedraagt voor de onderzochte setting, bevat de dataset geen annotaties van de grondwaarheid die empirische validatie van hun nauwkeurigheid mogelijk maken. Belangrijk is dat onderzoeken wijzen op de behoefte aan gespecialiseerde datasets die

dergelijke modellering vergemakkelijken en inzicht verschaffen in de uitdagingen die moeten worden overwonnen om deze te creëren.

Over het algemeen wijzen de bevindingen van het proefschrift in de richting van het belang van het beschouwen van persoonlijke herinneringen in automatische effectvoorspellingen en benadrukken ze de potentiële voordelen hiervan in bestaande technologische systemen. Bovendien biedt de ontwikkelde Artificial Empathic Memory-architectuur een blauwdruk die toekomstig onderzoek over het onderwerp kan structureren en sturen. Meer in het algemeen bieden onze studies die persoonlijke herinneringen combineren met andere contextuele variabelen empirisch bewijs voor de algehele waarde van contextgevoelige benaderingen om de nauwkeurigheid en robuustheid van automatische effectvoorspelling te verbeteren.

ACKNOWLEDGEMENTS

This chapter marks the final step on my journey as a Ph.D. candidate, and I could not have completed it without the help and support of a great many people.

I want to thank the members of my defense committee – Dr. Alan Hanjalic, Dr. Mohammad Soleymani, Dr. Martha Larson, Dr. Dirk Heylen, and Dr. Marcel Reinders – for their participation in this process and the opportunity to discuss my research project with them. I am especially deeply indebted to my team of promoters: without your knowledge, passion, and patience, it would have been impossible for me to complete this project. Each of you helped to shape this project (and me as a person) differently over the past years.

To Mark: Your seemingly unwavering trust and kind calmness throughout all phases of this project – no matter how hectic or challenging the circumstances – were truly inspiring. More than once, you dissolved some conceptual knot that I may have been struggling with for weeks with just a casual comment in one of our meetings. Thank you for believing in me and my ideas! This project would not be the same without your guidance.

To Hayley: It isn't easy to express the many ways in which you have contributed to both my research and personal growth. During the first year of my project, I was often lost and riddled with self-doubt. However, from the onset, you have challenged me to go outside of my comfort zone, push myself further, and carve a spot for the work that I believe in into the existing research landscape. More than anything else, this helped me stop doubting and start believing in myself and the value of my ideas. Over the years, your sheer enthusiasm and drive for doing research were incredible to behold. I hope that just a sliver of it has rubbed off on me. I may never manage to stop writing my articles like a sociologist from the 1960s, but when I do, every word I end with an "s' "will make me smile with fond memories. Thank you!

To Joost: I remember being quite skeptical of this "Affective Computing" topic when I first stepped into your seminar in the Winter of 2013. However, you quickly and fundamentally changed my opinion. I cannot imagine having undertaken this Ph.D. journey without you being involved. Your passion for understanding the human mind and the incredible intellectual and methodological breadth with which you approach this subject has been a constant source of inspiration. You make research in all its stages feel exciting and fun, and that talent of yours was invaluable throughout this project. I hope to carry a bit of your vigor and sense of wonder along with me. Thank you!

Finally, I also want to thank Marieke Peeters: you supervised me only for a short period at the beginning of this journey, but your guidance and support in these early days resonated throughout the remaining years. Thank you!

I had the privilege to engage in many insightful discussions with my colleagues at the Interactive Intelligence group (both past and present). I am very grateful for that:

Aleksander, Anita, Birna, Carolina, Catha, Catholijn, Chris, Elena, Elie, Emma, Enrico, Frans, Ilir, Jasper, Joachim, Joana, Koen, Luciano, Masha, Merijn, Michiel, Miguel, Mike, Myrthe, Pradeep, Rijk, Roel, Rolf, Ursula, Vincent, and Wouter. I want to especially thank Willem-Paul (for all the advice and interesting pointers he gave me over the years), and my long-time office mates for their lovely company: Ding (the Secret Mastermind), Franziska (the Heart and Soul of II), Frank (the Heart and Soul of Zoetermeer), Thomas (the Riddler), and Rifca (with the kindest heart).

Similarly, I want to thank all the members of the Pattern Recognition and Bio Informatics group that I had the pleasure of interacting with or even just listen to in a coffee talk: Amogh, Arman, Attila, Bob, Burak, David, Hadi, Jan, Jesse, Jin, Marco, Nergis, Ombretta, Osman, Ramin, Robert-Jan, Saskia, Seyran, Skander, Soufiane, Stephan, Taylan, Tom, Yancong, Yanxia, Yeshwanth (Niece was a blast – I don't think I've ever laughed so hard in my life!), Yuko, and Ziqi. Being surrounded by so many incredibly knowledgeable and talented people over the years was immensely stimulating and has provided me with many different perspectives.

Of course, I am especially indebted to my fellow "BlaBla"-ers at the Socially Perceptive Computing group (again both present and past): Chirag (the Critical Woodchucker), Jose (the Conversational Ninja), Stephanie (the No-nonsense Superhero), Ekin and Laura (the Hardened Veterans).

In addition, many a day was carried by the incredible technical support provided by Ruud, Bart, and Robbert. Thank you!

Throughout the project, I had the great pleasure to exchange ideas and collaborate with the other members of the 4TU Human & Technology consortium. I'm especially grateful for the time I got to spend with the other PhD students there: Jan, Minha, Michel (the Chef – let's never go to Upstate New York, ok?), and Sima.

A special thanks also to former teachers and mentors Daniel Fetzner, Maarten Laamers, and Peter van der Putten. Without your encouragement, I would never have discovered my passion for research.

Of course, I would not have managed to complete this journey without the support of many friends along the way: Pietro ("Just go and get a Mellon. It will be fine"), Frank (again), Malte (Mr. Crossbow Regulations), Berend (the only true Dragon Lord I know), and Rijk (you gave me veganism, damn you); Jose and Michel-Pierre (again); Andres and Elisa (Little Italy will always be in my heart); Wouter, Remi, Alice, Joanna, Jorrit, Dominik, Ronny, Patrick, and Sandro (memories of good old times is what this is all about, after all); Michelle and Alicia (when will we celebrate again?).

Finally, no one deserves greater thanks than the members of my family, who, through their many sacrifices and support over the last five years, have enabled me to complete this project.

To Sara: When I started this project, we had just gotten married, and you left everything and everyone behind to follow me. Without your love, support, and insight, I would likely have given up a hundred times over, but together we've pulled through. I owe you everything, and I love you with all my heart.

To my mother: Ich weiß, dass du für mich im Verlauf meines Lebens viel erduldet und getan hast. Besonders in den letzten Jahren war es bestimmt oft nicht einfach. Ohne dich, hätte ich dieses Projekt niemals beginnen können, geschweige denn es zuendebringen. Ich danke dir von ganzem Herzen und mit all meiner Liebe dafür.

To Shamsi: I know you're just a cat and can't read (yet), but in the darkest hours before the harshest deadlines – when everybody else was asleep –, you were always there to offer up emotional support. Much love and thanks for that!

I am sure that I have forgotten to mention at least one person on this (already long) list. I hope you can forgive me. I truly appreciate the help and support you have given me and offer you a big: "Thank you!".

CURRICULUM VITÆ

Bernd Dudzik holds a research position in the Intelligent Systems department at the Delft University of Technology. He has obtained a Bachelor's degree in Online Media at Furtwangen University (Germany) and a Master's degree in Media Technology at Leiden University (The Netherlands). His research is situated in the areas of Affective Computing and Social Signal Processing. It focuses on enabling technological systems to infer cognitive and affective states from data about human behavioral cues using multimodal machine learning. He is particularly interested in developing context-sensitive approaches for this kind of computer-based behavioral analysis. By accounting for person- and situation-specific influences on the underlying psychological processes driving human behavior in automatic inferences, he strives to increase their accuracy and robustness across different scenarios in the real world.
Bernd is a member of the *Association for the Advancement of Affective Computing (AAAC)*, as well as a regular reviewer for the *IEEE Transactions on Affective Computing (TAFFC)* and the *International Conference on Affective Computing & Intelligent Interaction (ACII)*. Beyond his research activities, Bernd has been involved in interdisciplinary media art projects that bridge technology, art, and design. They have been showcased in Germany, Egypt, and The Netherlands.

List of Related Publications

1. **Dudzik, B.**, Neerincx, M., Hung, H., & Broekens, J. (2018). *Artificial empathic memory: Enabling media technologies to better understand subjective user experience.* EE-USAD 2018 - Proceedings of the 2018 Workshop on Understanding Subjective Attributes of Data, with the Focus on Evoked Emotions, Co-Located with MM 2018. https://doi.org/10.1145/3267799.3267801

2. **Dudzik, B.**, Hung, H., Neerincx, M. A., & Broekens, J. (2021). *Collecting Mementos: A Multimodal Dataset for Context-Sensitive Modeling of Affect and Memory Processing in Responses to Videos.* IEEE Transactions on Affective Computing, 3045(c), 1–1. https://doi.org/10.1109/TAFFC.2021.3089584

3. **Dudzik, B.**, Hung, H., Neerincx, M., & Broekens, J. (2020). *Investigating the Influence of Personal Memories on Video-Induced Emotions.* Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 53–61. https://doi.org/10.1145/3340631.3394842

4. **Dudzik, B.**, Broekens, J., Neerincx, M., & Hung, H. (2020). A Blast From the Past: Personalizing Predictions of Video-Induced Emotions using Personal Memories as Context. ArXiv. https://arxiv.org/abs//2008.12096

5. **Dudzik, B.**, Broekens, J., Neerincx, M., & Hung, H. (2020). *Exploring Personal Memories and Video Content as Context for Facial Behavior in Predictions of Video-Induced Emotions.* Proceedings of the 2020 International Conference on Multimodal Interaction, 10(20), 153–162. https://doi.org/10.1145/3382507.3418814

6. **Dudzik, B.**, Olenick, J., Broekens, J., Chang, C. H., Hung, H., Neerincx, M., & Kozlowski, S. W. J. (2018). *Discovering digital representations for remembered episodes from lifelog data.* Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, MCPMD 2018. https://doi.org/10.1145/3279810.3279850