# Exploring Personal Memories and Video Content as Context for Facial Behavior in Predictions of Video-Induced Emotions

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

**DOI**
10.1145/3382507.3418814

**Publication date**
2020

**Document Version**
Final published version

**Published in**
ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction

# Exploring Personal Memories and Video Content as Context for Facial Behavior in Predictions of Video-Induced Emotions

Bernd Dudzik*
Delft University of Technology
Delft, South Holland, The Netherlands
B.J.W.Dudzik@tudelft.nl

Joost Broekens
Leiden University
Leiden, South Holland, The Netherlands
D.J.Broekens@tudelft.nl

Mark Neerincx
Delft University of Technology
Delft, South Holland, The Netherlands
M.A.Neerincx@tudelft.nl

Hayley Hung
Delft University of Technology
Delft, South Holland, The Netherlands
H.Hung@tudelft.nl

## 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 article, 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.

## CCS CONCEPTS

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

## KEYWORDS

Emotion Recognition; Affect Detection; Context-Awareness

**ACM Reference Format:**
Bernd Dudzik, Joost Broekens, Mark Neerincx, and Hayley Hung. 2020. 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 (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418814

*This is the corresponding author


*ICMI '20, October 25–29, 2020, Virtual event, Netherlands*

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418814

## 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 [4]. 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 [30].

The automatic analysis of facial behavior is traditionally an essential method for automatic affect detection [19], including the recognition of emotional responses to video stimuli (e.g., [44, 61, 62]). 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 [26]. 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 [3]. 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 [32], as well as the perceptions of third-party observers [25].

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 [31, 42, 67]. 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., [41]). However, context-sensitive approaches remain underexplored in automatic affect detection [23], despite researchers generally acknowledging their potential [60, 66, 68]. 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 [29]. Overcoming these challenges requires systematic exploration of person- and situation-specific influences in computational modeling activities [23] informed by findings from the social sciences [3]. 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., [62]). However, numerous other important influences exist [60].

In this article, 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 [5, 43] and that the evoked memories are a powerful causal influence on emotional responses [35]. Moreover, the feelings associated with any memories triggered by a video in this way closely relate to its overall emotional impact [22], 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 [55, 65], for example, in service of social bonding or emotion regulation processes [7]. There is evidence that people share memories for similar reasons on social media [11], and that they readily describe memories triggered in them by social media content [17]. Additionally, research is extensively exploring both the automatic affective analysis of text-data from social media [47], and that of face-to-face dialog [12]. 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 [20]. 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 article, 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.

## 2 BACKGROUND AND RELATED WORK

### 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 [31]. A basic form of sensory information is other behavioral signals and cues, e.g., body posture and gestures [67]. Such *cross-behavior context* has been extensively explored in multimodal analysis approaches, especially with speech as an added modality, typically showing performance improvements [19]. 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. [56] 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 [67]. 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 [42]. 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 [23]. Notably, however, Kosti et al. [38] 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., [36, 44, 62]). Similarly, when viewing occurs in a social setting with multiple persons, looking at other viewers' behavior might provide context for predictions in computational models [46].

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.

### 2.2 Representing Affective States for Detection

A challenging aspect of developing systems for automatic affect detection is the conceptualization of the targeted states [19], 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 [3]. 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 [57]. A widely used dimensional framework is *Pleasure-Arousal-Dominance (PAD)* [45]. 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 [39]) and automatic affect detection (e.g. *DEAP* [36], or *EMOTIC* [37] 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 [9].

## 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.

### 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 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 1*).

**Table 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 |

### 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 [36]. We chose these stimuli for two reasons: (1) the strong capacity of music to trigger personal memories [35], 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.

### 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 [10] 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 article. This filtering left us with a set of 932 recordings of viewers' responses for which both memory descriptions and behavior are available.

# 4 PREDICTIVE MODELING

## 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 [19]. 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. [66], and more recently Zhao et al. [68]). 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 [19]. 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 usi

ng stacked generalization on this task, compared to feature-level fusion [20]. Motivated by this, we conduct all our experiments using only this approach to decision-level fusion. *Figure 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-sets that we extract from them comprising different modalities for fusion and predictions can be found in *Table 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 [20]. 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* [53].

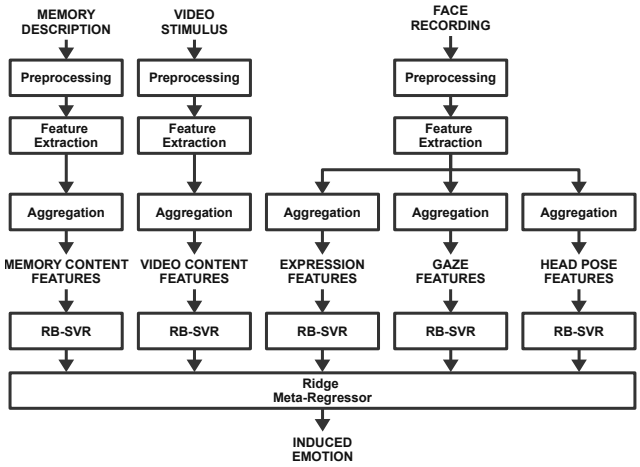


**Figure 1: Overview of our approach for predictive modeling and decision-level multimodal fusion.**

**Table 2: Overview of Extracted and Modality-specific Feature Sets from Input Sources and their Aggregation**

| MODALITY | FEATURES | # EXTR. | SOURCE | # AGGR. |
|---|---|---|---|---|
| **E** | Action Units | 17 | Face Rec. | 13498 |
| **G** | Direction | 8 | Face Rec. | 6352 |
| **P** | Pos./Orient. | 6 | Face Rec. | 4764 |
| **V** | Theory-inspired | 271 | Vid. Stim. | 271 |
| | Deep Visual | 4096 | | 4096 |
| | Visual Sentiment | 4342 | | 4342 |
| | openSMILE | 1582 | | 1582 |
| **M** | Lexical | 130 | Mem. Descr. | 130 |
| | W. Embeddings | 500 | | 500 |

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

## 4.2 Face Recordings Processing

We deploy the software *OpenFace 2.0* [2] 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 [16], which implements 63 best practice methods for time series characterization, computing a total of 794 generic features [1] per series. See *Table 2* for details about the amount of extracted and aggregated features per response.

[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

***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.

***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.

## 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 [40], as well as those of Bhattacharya et al. [6] to characterize each of the extracted video frames. This combination has been used previously in affective content analysis (e.g., [58]).

***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 [59] from the Keras framework for python [15] 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 [18].

***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., [44, 58]). 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 [13] 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 [57] for a detailed description). Benchmarking challenges for affective content analysis have used these features as a baseline approach [18].

## 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 [47]): (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 [1, 8, 14, 33, 48–52, 63, 64] has achieved state-of-the-art performance for affect regression [24]. We extended this list by a new source containing word-level ratings for Pleasure, Arousal, and Dominance [47], and lexica-based VADER Sentiment ratings [34]. 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 [54] 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 [69].

# 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

**Table 3: 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;

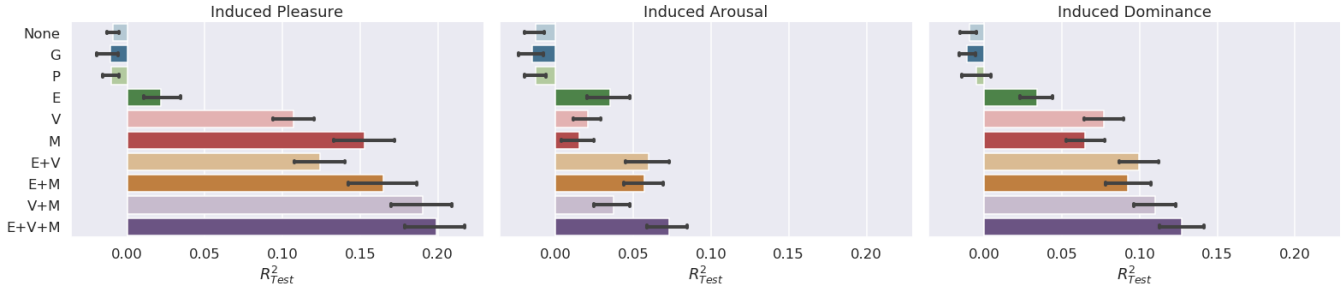


**Figure 2: Test-performance ($R^2_{Test}$) of our model for Induced Pleasure, Arousal and Dominance when using Facial Expressions (*E*), Gaze (*G*), Head Pose (*P*), Video Content (*V*), Memory Content (*M*), or their multimodal fusions for predictions. *None* is a baseline always predicting the mean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

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?).

## 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.

## 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 2*. Furthermore, *Table 3* 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

[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. [27] 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 3.

**Table 4: 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;

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 provides substantial benefits to facial analysis for predicting pleasure and dominance. The same is true for access to video content, either alone or in combination with memory content. However, neither context modality facilitates improvements over facial expressions for arousal.

***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 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 4* 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 ($M$vs$V$: $t(29) = 2.17$,$p < .05$), or facial expressions ($M$vs$E$: $t(29) = 9.37$, $p < .001$). 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.

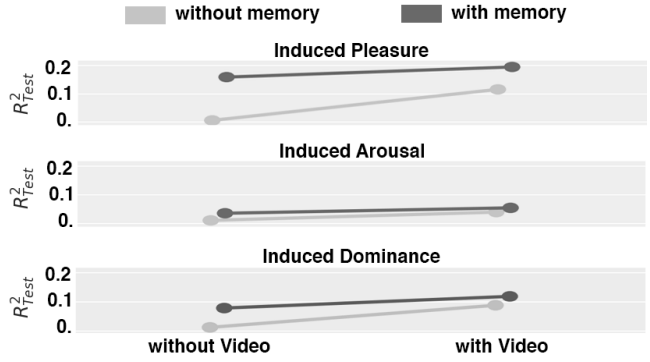


**Figure 3: Marginal means of Model Performance ($R^2_{Test}$) with/without access to Memory and Video Content modalities. Converging lines indicate negative interactions due to overlapping information.**

Further, inspection reveals no significant interactions between the context modalities and facial expressions ($E$), 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 3* 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.

## 6 DISCUSSION

***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 significant performance benefits 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 [20]. 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 [22], and accounting for their impact in automatic predictions could facilitate a broad range of novel applications [21], 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 provides in isolation is comparatively low. This observation is congruent with the findings of Hirt et al. [32], 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 [28], 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.

***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 [56]. 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.

## 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 article 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.

## ACKNOWLEDGMENTS

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

## REFERENCES


[1] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.. In *Lrec*, Vol. 10. 2200–2204.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66. https://doi.org/10.1109/FG.2018.00019

[3] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest* 20, 1 (jul 2019), 1–68. https://doi.org/10.1177/1529100619832930

[4] Anne Bartsch. 2012. Emotional Gratification in Entertainment Experience. Why Viewers of Movies and Television Series Find it Rewarding to Experience Emotions. *Media Psychology* 15, 3 (jul 2012), 267–302. https://doi.org/10.1080/15213269.2012.693811

[5] Amy M. Belfi, Brett Karlan, and Daniel Tranel. 2016. Music evokes vivid autobiographical memories. *Memory* 24, 7 (aug 2016), 979–989. https://doi.org/10.1080/09658211.2015.1061012

[6] Subhabrata Bhattacharya, Behnaz Nojavanasghari, Tao Chen, Dong Liu, Shih-Fu Chang, and Mubarak Shah. 2013. Towards a comprehensive computational model foraesthetic assessment of videos. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*. ACM Press, New York, New York, USA, 361–364. https://doi.org/10.1145/2502081.2508119

[7] Susan Bluck, Nicole Alea, Tilmann Habermas, and David C. Rubin. 2005. A TALE of Three Functions: The Self–Reported Uses of Autobiographical Memory. *Social Cognition* 23, 1 (feb 2005), 91–117. https://doi.org/10.1521/soco.23.1.91.59198

[8] Felipe Bravo-Marquez, Eibe Frank, Saif M. Mohammad, and Bernhard Pfahringer. 2016. Determining Word-Emotion Associations from Tweets by Multi-label Classification. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 536–539. https://doi.org/10.1109/WI.2016.0091

[9] Joost Broekens. 2012. In Defense of Dominance. *International Journal of Synthetic Emotions* 3, 1 (jan 2012), 33–42. https://doi.org/10.4018/jse.2012010103

[10] Joost Broekens and Willem-Paul Brinkman. 2013. AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies* 71, 6 (jun 2013), 641–667. https://doi.org/10.1016/j.ijhcs.2013.02.003

[11] Barbara Caci, Maurizio Cardaci, and Silvana Miceli. 2019. Autobiographical memory, personality, and Facebook mementos. *Europe's Journal of Psychology* 15, 3 (sep 2019), 614–636. https://doi.org/10.5964/ejop.v15i3.1713

[12] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. 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, 39–48. https://doi.org/10.18653/v1/S19-2005

[13] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. (oct 2014). arXiv:1410.8586

[14] Yoonjung Choi and Janyce Wiebe. 2014. +/-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, 1181–1191. https://doi.org/10.3115/v1/D14-1125

[15] François Chollet and Others. 2015. Keras. \url{https://keras.io}.

[16] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. 2018. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* 307 (sep 2018), 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

[17] Dan Cosley, Victoria Schwanda Sosik, Johnathon Schultz, S Tejaswi Peesapati, Soyoung Lee, Tejaswi Peesapati, and Soyoung Lee. 2012. Experiences With Designing Tools for Everyday Reminiscing. *Human–Computer Interaction* Volume 27, July (2012), 175–198. https://doi.org/10.1080/07370024.2012.656047

[18] Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, Zhongzhe Xiao, and Mats Sjöberg. [n.d.]. The MediaEval 2018 emotional impact of Movies task. *CEUR Workshop Proceedings* ([n. d.]), 1–3.

[19] Sidney K. D'mello and Jacqueline Kory. 2015. A Review and Meta-Analysis of Multimodal Affect Detection Systems. *Comput. Surveys* 47, 3 (apr 2015), 1–36. https://doi.org/10.1145/2682899

[20] Bernd Dudzik, Joost Broekens, Mark Neerincx, and Hayley Hung. 2020. A Blast From the Past: Personalizing Predictions of Video-Induced Emotions using Personal Memories as Context. arXiv:2008.12096

[21] Bernd Dudzik, Hayley Hung, Mark Neerincx, and Joost Broekens. 2018. 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, 1–8. https://doi.org/10.1145/3267799.3267801

[22] Bernd Dudzik, Hayley Hung, Mark Neerincx, and Joost Broekens. 2020. 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, 53–61. https://doi.org/10.1145/3340631.3394842

[23] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk K.J. Heylen, Hayley Hung, Mark A. Neerincx, and Khiet P. Truong. 2019. 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, 206–212. https://doi.org/10.1109/ACII.2019.8925446

[24] Venkatesh Duppada, Royal Jain, and Sushant Hiray. 2018. SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets. *Proceedings of The 12th International Workshop on Semantic Evaluation* (apr 2018), 18–23. arXiv:1804.06137

[25] Damien Dupré, Eva G. Krumhuber, Dennis Küster, and Gary J. McKeown. 2020. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLOS ONE* 15, 4 (apr 2020), e0231968. https://doi.org/10.1371/journal.pone.0231968

[26] José Miguel Fernández-Dols and Carlos Crivelli. 2013. Emotion and expression: Naturalistic studies. *Emotion Review* 5, 1 (2013), 24–29. https://doi.org/10.1177/1754073912457229

[27] Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R*. Sage publications.

[28] Agneta H. Fischer and Antony S. R. Manstead. 2008. Social functions of emotion. In *Handbook of emotions (3rd ed.)* (3 ed.), M Lewis, J M Haviland-Jones, and L F Barrett (Eds.). Guilford Press, New York, NY, US, Chapter 28, 456–468.

[29] Zakia Hammal and Merlin Teodosia Suarez. 2015. 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, 1–2. https://doi.org/10.1109/FG.2015.7284841

[30] Alan Hanjalic and Li-Qun Xu. 2005. Affective video content representation and modeling. *IEEE Transactions on Multimedia* 7, 1 (feb 2005), 143–154. https://doi.org/10.1109/TMM.2004.840618

[31] Ursula Hess and Shlomo Hareli. 2015. 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, 1–6. https://doi.org/10.1109/FG.2015.7284842

[32] Franziska Hirt, Egon Werlen, Ivan Moser, and Per Bergamin. 2019. Measuring emotions during learning: lack of coherence between automated facial emotion recognition and emotional experience. *Open Computer Science* 9, 1 (jan 2019), 308–317. https://doi.org/10.1515/comp-2019-0020

[33] Minqing Hu and Bing Liu. 2004. 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, 168. https://doi.org/10.1145/1014052.1014073

[34] C J Hutto and Eric Gilbert. 2014. 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*. 216–225.

[35] Petr Janata, Stefan T. Tomic, and Sonja K. Rakowski. 2007. Characterisation of music-evoked autobiographical memories. *Memory* 15, 8 (nov 2007), 845–860. https://doi.org/10.1080/09658210701734593

[36] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (jan 2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[37] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2017. EMOTIC: Emotions in Context Dataset. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Vol. 2017-July. 2309–2317. https://doi.org/10.1109/CVPRW.2017.285

[38] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotion Recognition in Context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1960–1968. https://doi.org/10.1109/CVPR.2017.212

[39] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, and Others. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* 1 (1997), 39–58.

[40] Jana Machajdik and Allan Hanbury. 2010. 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, 83. https://doi.org/10.1145/1873951.1873965

[41] Andreas Marpaung and Avelino Gonzalez. 2017. 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, 454–467. https://doi.org/10.1007/978-3-319-57837-8_38

[42] David Matsumoto and Hyi Sung Hwang. 2010. Judging Faces in Context. *Social and Personality Psychology Compass* 4, 6 (jun 2010), 393–402. https://doi.org/10.1111/j.1751-9004.2010.00271.x

[43] Daniel G. McDonald, Melanie A. Sarge, Shu-Fang Lin, James G. Collier, and Bridget Potocki. 2015. A Role for the Self: Media Content as Triggers for Involuntary Autobiographical Memories. *Communication Research* 42, 1 (feb 2015), 3–29. https://doi.org/10.1177/0093650212464771

[44] Daniel McDuff and Mohammad Soleymani. 2017. 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, 339–345. https://doi.org/10.1109/FG.2017.49

[45] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology* 14, 4 (dec 1996), 261–292. https://doi.org/10.1007/BF02686918

[46] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2017. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *Expert Systems with Applications* 39, 16 (feb 2017), 12378–12388. https://doi.org/10.1016/j.eswa.2012.04.084 arXiv:1702.02510

[47] Saif Mohammad. 2018. 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)*. 174–184.

[48] Saif Mohammad and Peter Turney. 2010. 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, 26–34.

[49] Saif M. Mohammad. 2017. Word Affect Intensities. *LREC 2018 - 11th International Conference on Language Resources and Evaluation* (apr 2017), 174–183. arXiv:1704.08798 http://arxiv.org/abs/1704.08798

[50] Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence* 31, 2 (may 2015), 301–326. https://doi.org/10.1111/coin.12024

[51] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. **SEM 2013 - 2nd Joint Conference on Lexical and Computational Semantics* 2 (aug 2013), 321–327. arXiv:1308.6242

[52] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CEUR Workshop Proceedings* 718 (mar 2011), 93–98. arXiv:1103.2903

[53] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and Others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[54] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. 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, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[55] Bernard Rimé, Susanna Corsini, and Gwénola Herbette. 2002. Emotion, verbal expression, and the social sharing of emotion. *The verbal communication of emotions: Interdisciplinary perspectives* (2002), 185–208.

[56] Philipp V. Rouast, Marc Adam, and Raymond Chiong. 2018. Deep Learning for Human Affect Recognition: Insights and New Developments. *IEEE Transactions on Affective Computing* (2018). https://doi.org/10.1109/TAFFC.2018.2890471

[57] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association*.

[58] Michael James Scott, Sharath Chandra Guntuku, Weisi Lin, and Gheorghita Ghinea. 2016. Do Personality and Culture Influence Perceived Video Quality and Enjoyment? *IEEE Transactions on Multimedia* 18, 9 (sep 2016), 1796–1807. https://doi.org/10.1109/TMM.2016.2574623

[59] Karen Simonyan and Andrew Zisserman. 2015. 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. arXiv:1409.1556

[60] Mohammad Soleymani, Martha Larson, Thierry Pun, and Alan Hanjalic. 2014. Corpus Development for Affective Video Indexing. *IEEE Transactions on Multimedia* 16, 4 (jun 2014), 1075–1089. https://doi.org/10.1109/TMM.2014.2305573 arXiv:1211.5492

[61] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (jan 2012), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

[62] Jennifer J. Sun, Ting Liu, Alan S. Cowen, Florian Schroff, Hartwig Adam, and Gautam Prasad. 2020. EEV Dataset: Predicting Expressions Evoked by Diverse Videos. (jan 2020). arXiv:2001.05488 http://arxiv.org/abs/2001.05488

[63] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. , 24–54 pages. https://doi.org/10.1177/0261927X09351676

[64] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2544–2558.

[65] W. Richard Walker, John J. Skowronski, Jeffrey A. Gibbons, Rodney J. Vogl, and Timothy D. Ritchie. 2009. Why people rehearse their memories: Frequency of use and relations to the intensity of emotions associated with autobiographical memories. *Memory* 17, 7 (oct 2009), 760–773. https://doi.org/10.1080/09658210903107846

[66] Shangfei Wang and Qiang Ji. 2015. Video Affective Content Analysis: A Survey of State-of-the-Art Methods. *IEEE Transactions on Affective Computing* 6, 4 (oct 2015), 410–430. https://doi.org/10.1109/TAFFC.2015.2432791

[67] Matthias J. Wieser and Tobias Brosch. 2012. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology* 3, NOV (nov 2012), 471. https://doi.org/10.3389/fpsyg.2012.00471

[68] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. 2020. Affective Computing for Large-scale Heterogeneous Multimedia Data. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 3s (jan 2020), 1–32. https://doi.org/10.1145/3363560 arXiv:1911.05609

[69] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.