Exploring the Detection of Spontaneous Recollections during Video-viewing In-the-Wild using Facial Behavior Analysis

Dudzik, Bernd; Hung, Hayley

DOI
10.1145/3536221.3556609

Publication date
2022

Document Version
Final published version

Published in
ICMI 2022 - Proceedings of the 2022 International Conference on Multimodal Interaction

Exploring the Detection of Spontaneous Recollections during Video-viewing In-the-Wild using Facial Behavior Analysis

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

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

ABSTRACT

Intelligent systems might benefit from automatically detecting when a stimulus has triggered a user's recollection of personal memories, e.g., to identify that a piece of media content holds personal significance for them. While computational research has demonstrated the potential to identify related states based on facial behavior (e.g., mind-wandering), the automatic detection of spontaneous recollections specifically has not been investigated this far. Motivated by this, we present machine learning experiments exploring the feasibility of detecting whether a video clip has triggered personal memories in a viewer based on the analysis of their Head Rotation, Head Position, Eye Gaze, and Facial Expressions. Concretely, we introduce an approach for automatic detection and evaluate its potential for (1) person-independent and (2) video-independent predictions using in-the-wild webcam recordings. Overall, our findings demonstrate the capacity for above chance detections in both settings, with substantially better performance for the video-independent variant. Beyond this, we investigate the role of person-specific recollection biases for predictions of our video-independent models and the importance of specific modalities of facial behavior. Finally, we discuss the implications of our findings for detecting recollections and user-modeling in adaptive systems.

CCS CONCEPTS

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

KEYWORDS

User-Modeling, Affective Computing, Recollection, Memories, Mind-Wandering, Cognitive Processing, Facial Behavior Analysis

ACM Reference Format:
Bernd Dudzik and Hayley Hung. 2022. Exploring the Detection of Spontaneous Recollections during Video-viewing In-the-Wild using Facial Behavior Analysis. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3536221.3556609

*This is the corresponding author

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11.
https://doi.org/10.1145/3536221.3556609

1 INTRODUCTION

The personal memories spontaneously triggered by objects, places, and people have a substantial influence in defining their meaning to us, e.g., as treasured mementos or beloved friends. Importantly, the memories that stimuli trigger in us are closely connected to our emotional responses and attitudes to them [15]. For example, a photograph that reminds us of a trip with good friends might lift our mood when we look at it, holding deep sentimental value. Similarly, visiting a location where someone had an accident in the past might bring that incident to mind and shape their behavior during the visit. As such, technologies that strive to provide socio-emotional support to people throughout their daily lives could benefit from understanding when memories occur for a broad range of purposes [14]. For example, a personal media system might use detections of occurring recollections to learn whether a particular piece of content is evocative to a user without requiring self-reports. Such evocative stimuli could then be drawn on later for relevant recommendations in applications (e.g., to stimulate memories in dementia treatment [29]).

While spontaneous recollections are common, psychological research has only begun to unravel the conditions under which situations are conducive for involuntary memories to be triggered in a person [4, 5, 36, 37]. As such, computational modeling of recollections for use in intelligent systems has mainly been limited to the area of cognitive agents to synthesize plausible interaction behavior [25]) but has not been addressed in user-modeling research. However, experiencing recollections can also be considered a state where attention is primarily internally directed [8]. In such states, an individual's focus is not absorbed by some external event in the world (e.g., a beautiful view) but instead by cognitive events inside their mind (e.g., when daydreaming) [8]. Work on detecting such attentional states for user models exists, and it has typically targeted the umbrella phenomenon of mind-wandering (e.g., Bixler et al. [6]). Findings here have demonstrated successful detections from facial behavior (e.g., Steward et al. [35]). While the majority of such studies have taken place in the laboratory to ensure experimental control (e.g., Annerer-Walcher et al. [1]), some studies have demonstrated potential for predictions in the wild (e.g., a classroom setting [7, 30]). Particularly eye gaze behaviors have been identified as indicative of internally-directed attention [1, 22]. However, no research in this area has specifically explored the automatic detection of spontaneous recollections. Nevertheless, their findings lend credibility to the assumption that detections based on facial behavior are at least principally possible.

This paper describes our attempts to empirically explore the feasibility of automatically detecting the occurrence of recollections

based on visual recordings of a person's facial behavior. Our investigation focuses on detecting memories triggered in a person when they watch a video clip in an in-the-wild scenario. We target this particular setting since it is a good initial situation for exploration. First, it provides a controlled exposure to triggering stimuli compared to many other real-world situations (e.g., social interactions). In particular, a video's content is fixed and likely forms a primary target of a viewer's visual attention. Additionally, video-viewing offers many aspects of ecological validity (i.e., exposure to information from multiple senses and semantically meaningful content, potentially with a personal relevance). For this reason, videos are generally popular for elicitation of complex and dynamic psychological phenomena, such as emotions [19], and are potent triggers for spontaneous recollections [26]. Crucially, estimating the evocative properties of videos might be of immediate interest for a range of applications. For example, it could aid in recommending evocative content on social media for entertainment [9] or well-being [34], and support reminiscence therapy with dementia patients [27]. Finally, with the *Mementos* dataset [16] a resource capturing memories evoked by videos and associated behavioral responses already exists and is publicly available for study. This availability makes video-triggered memories a convenient starting point for our explorations. Overall, we present the following contributions:

- A technological architecture for automatic facial behavior analysis to detect the occurrence of recollections. This method is intended as suitable proof of concept to explore this novel detection task's feasibility empirically. It uses frame-wise extraction of facial behavior features from video signals in combination with the *MINImally RandOm Convolutional KErnel Transformation (MINIROCKET)* algorithm [10] for time series representation.
- An empirical exploration of the feasibility of detecting spontaneous recollections during video-viewing from facial behavior in the wild. Concretely, we investigate (1) *person-independent* and (2) *video-independent* predictions in a series of machine learning experiments. These demonstrate the above random performance for our approach in both settings. However, our findings indicate substantial accuracy benefits for video-independent predictions.
- A detailed analyses of (1) the role played by person-specific recollection biases for predictions of our models, as well as (2) the contribution of specific modalities (e.g., Head Rotation or Eye Gaze) for their performance.

2 BACKGROUND AND RELATED WORK

The link between memory recollection and facial behavior has only been tentatively explored in existing psychological research [17, 32]. However, these findings indicate a connection between the dispersion of fixations in eye gaze behavior [32]), as well as differences in the amount of emotional facial expressions between episodic and semantic recollections of past experiences [17]. No work on automatic behavioral analysis or user-modeling up to this point has specifically explored the detection of spontaneous recollections, i.e., memories that are involuntarily triggered by stimuli. Nevertheless, a recent study by Gupta et al. [20] demonstrates that measures of autonomic activity and eye gaze can automatically distinguish between conditions in which participants are explicitly instructed to recall emotional autobiographical memories and a neutral control state. While these findings stem from a highly constrained setting in a Virtual Reality environment and a small number of participants (i.e., $N = 20$), they nevertheless show a promising connection between gaze behavior and another type of recollection process (i.e., voluntary or intentional recall).

More generally, being immersed in the experience of a personal memory can also be considered a state of internally-directed attention (see Christoff et al. [8] for a relevant taxonomy). In such states, the focus of a person's attention is not (or no longer) focused on an external task or stimulus (e.g., the contents of a book while reading) but instead centers around mental content, such as thoughts or memories. Involuntary shifts from externally- towards internally-directed attentional focus are often studied under the umbrella term of *mind-wandering* [6]. Indeed, films have been identified as a kind of stimulus that results in mind-wandering episodes frequently revolving around the recollection of personal memories [18]. A substantial body of computational research has explored detecting such mind-wandering episodes from physiological [31] and particularly behavioral data [1, 7, 21, 22, 38]. Findings in such studies indicate that detections of internally-directed attention based on facial behavior – especially gaze-related aspects – are feasible but typically challenging [1, 7, 21, 35]. Conceptually most similar to our work are studies demonstrating the capacity of detecting mind-wandering or internally-directed attention from audiovisual data of facial behavior. For example, Steward et al. [35] explore detections during film comprehension, Bosch&D'Mello [7] during interactions with an intelligent agent, and Putze et al. [30] in a classroom setting. However, no existing work in this computational strain has focused specifically on detections of internally-directed attention concerning spontaneous recollection. Additionally, the majority of investigations undertaken thus far have focused on detections with a relatively high temporal resolution (i.e., segments of only several seconds in length [7, 22]) requiring elaborate self-report protocols (e.g., probing at specific intervals [7]). A primary motivation for this is the possibility for immediate intervention by intelligent systems when attention drifts away toward internal thoughts from important external tasks, such as driving [38] or studying [39]. Obtaining valid moment-to-moment self-reports for computational modeling in such settings is challenging. It is likely for this reason that existing studies often rely on a comparatively small number of participants [21, 22, 30, 31, 39] (However, notable exceptions exist, for example, the studies by Steward et al. [35] and Annerer-Walcher et al. [1]). Moreover, existing work on internally-directed attention often involves only a small degree of variation in stimuli to which responses are analyzed (e.g., during comprehension of only a single film [22, 35]). In contrast to these efforts, our work in this article focuses on detecting the occurrence of recollections at a global level of stimulus exposure instead of moment-to-moment shifts between internally- and externally-directed attention. Our goal is merely to determine whether any recollections have occurred during stimulus exposure or not, not exactly when. This task formulation is similar to *implicit tagging* of stimuli [33] (e.g., with emotion labels), potentially enabling downstream applications to act on tags describing a

video's evocative potential for a particular user, such as for retrieval or recommendation.

Another distinguishing feature of our investigation is that it draws on a dataset with a relatively large participant pool ($N = 298$) and broad range of different video stimuli ($N = 42$). Moreover, the Mementos dataset we used for analysis was captured in an in-the-wild setting instead of a laboratory. As such, the facial recordings are technologically challenging to analyze (e.g., varying lighting conditions or viewing angles; consumer-grade cameras) and ecologically valid (e.g., captured in the homes of participants where they naturally engage with video content).

3 DATASET

For our empirical investigations we rely on the recently released *Mementos* dataset [16]. This corpus has been specifically collected to support multimodal modeling of cognitive-affective memory processes, including spontaneous recollections. It captures the 1995 responses of 297 participants to a total of 42 different music videos and includes audiovisual recordings of participants' faces and self-reports about triggered memories. Below we provide a brief description of the protocol used to collect the dataset and its contents (for more details see the relevant publication by Dudzik et al. [16]).

3.1 Data Collection Procedure

The dataset was collected by exposing crowd-workers online to a selection of 42 music video segments from the DEAP corpus [23] and measuring their responses. Each participant was provided with 7 music videos from this pool. They were recruited until 50 responses had been collected for each video. During the exposure to a video, a viewer's face was recorded with the webcam embedded in their personal device. Once playback of the video finished, participants were provided self-reports about whether any memories had been triggered while viewing. If so, they had to provide additional descriptions of these recollections. The creators of the dataset filtered out several responses (e.g., due to technical corruption. This filtering leaves 1995 unique responses in the published corpus out of the original 2099 recorded samples. These samples are the responses of $N = 297$ unique participants ($N_{male} = 159$, $N_{female} = 138$). A more detailed description of the dataset can be found in its official publication [16].

3.2 Contents

3.2.1 Memory Self-Reports. After exposure to each video, participants in the Mementos dataset provide self-reports about whether any recollection was triggered in them while viewing or not. Out of the 1995 responses in the corpus, a total of 935 involves the recollection of at least one memory. Moreover, not all video clips shown to participants are equally evocative: some videos were substantially more likely to trigger memories in participants than others. Only 18% of viewers of the least evocative video experienced recollections, while the most evocative video resulted in recollections for up to 64% of the people exposed.

3.2.2 Face Recordings. Recordings were captured by participants' webcams when viewing the eliciting video clips online in their browser. Because of this, there is little control over the circumstances in which recordings took place, and they cover participants in many different viewing positions and lighting conditions. Moreover, due to latency issues, the duration of the face recordings in the Mementos dataset varies slightly between 50.33 and 69.86 ($M(SD) = 60.5(2.1)$) seconds. They are encoded at 30 frames per second at a 640×480 pixel resolution.

4 PREDICTIVE MODELING

4.1 Overview

This section describes our computational architecture for detecting spontaneous recollections from facial behavior. Its primary aim is to serve as a technologically valid and reproducible proof of concept for answering our research questions regarding the feasibility of detecting recollections in light of the available dataset. Because of the comparatively small size of the Mementos dataset, data efficiency is a concern and has been considered in selecting components throughout our architecture. Additionally, we are interested in exploring how aspects of facial behavior relate to the performance of predictions. For this reason, our approach relies on interpretable features (e.g., Facial Action Units) capturing aspects of facial activity as intermediate representations. In contrast to learned representations (i.e., Deep Learning), these intermediary features facilitate such investigations more easily through targeted ablation studies. For a graphical overview of our architecture, please have a look at *Figure 1*. At a high level, it comprises three distinct stages: (1) *Facial Behavior Represenation*, (2) *Time Series Represenation*, and (3) *Classification*. In our empirical investigations, we use two distinct versions of this overall architecture: RIDGE and PCA99-SVM (see Section 4.4 for details).

4.2 Facial Behavior Representation

To represent facial behavior in terms of features, we rely on *OpenFace 2.0* [3]. We extract both *Head Position*-features provided by OpenFace, i.e., in (x, y, z) world-coordinates, as well as the *Head Rotation*-features, i.e. $(pitch, yaw, rotation)$-angles relative to the camera. Additionally, we use the gaze-direction vectors provided by OpenFace to represent *Eye Gaze*. These are the angles in radians along the (x, y)-axis relative to the camera averaged across both eyes. Finally, we extract features relating to *Facial Expressions* in terms of the activation-intensity of 17 Facial Action Units (*AU Intensities*). These intensities range from $0 - 5$, with 0 denoting no activation and a value of 5 activation at maximum intensity. We combine the extracted frame-level features for each modality of facial behavior into time series by concatenating them along the time axis (the *Concat-Time* operation in *Figure 1*). This procedure results in a 25 univariate time series, one for each frame-level feature extracted via Openface. Because our approach for time series representation (see Section 4.3 below) requires data with a fixed length, we truncate each of these time series to that of the shortest series in our dataset (i.e., 50 seconds, or 1500 time steps). Finally, we replace time steps resulting from frames for which OpenFace provides low-confidence scores (i.e., $< .1$) with the mean value across the previous and following steps in a time series.

4.3 Time Series Representation

Representing time series in a manner that allows effective classification is a central component of research on bespoke Time Series

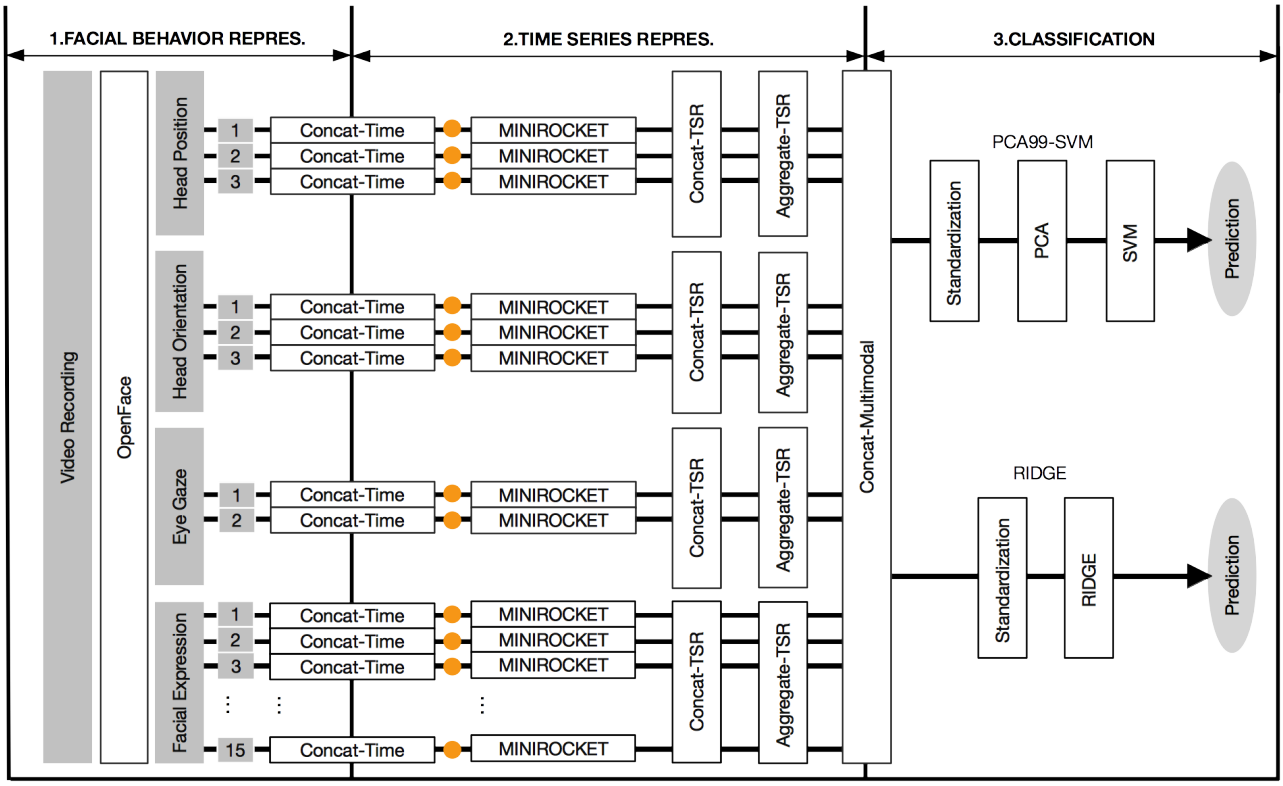

Figure 1: Overview of our Computational Architecture for Detecting the Occurrence of Spontaneous Recollections. All depicted components except for the Classification-part are shared between our two modeling approaches *RIDGE* and *PCAS99-SVM*. Orange marks indicate the position of the optional transformation operations for the experiments discussed in *Section 5.2.3*

Classification algorithms [2]. ROCKET (RandOm Convolutional KErnel Transformation) and its derivatives are a series of recently developed approaches for time series representation that facilitate state-of-the-art performance in a wide variety of tasks [10, 11]. The transformation algorithm comprises two principal steps: (1) applying a large amount (10000 or more) of convolutional kernels with random parameters to an input time series, and then (2) computing aggregate features over the feature map for each kernel (i.e., pooling), resulting in a high-dimensional representation.

In contrast to convolutional kernels in deep neural networks, the parameters for these kernels (e.g., their bias, length, and dilation) are not learned from the data but are sampled randomly. The initial ROCKET algorithm computes the *Maximum (max)* and *Proportion of Positive values (PPV)* for pooling of feature maps after convolution. The PPV pooling operation is a crucial element of ROCKET's performance and is calculated as the proportion of timesteps in a series for which the convolution output is positive [10]. Intuitively, this operation reflects the degree to which the pattern encoded by a particular kernel is present in the input data. MINIROCKET (MINImally-deterministic ROCKET) is an improved version of the basic ROCKET algorithm, minimizing stochastic behavior and computational demands without displaying reduced performance [11]. Instead of sampling kernel parameters at random, the algorithm uses a fixed set of 84-kernels, bias values drawn from the output of the convolution, as well as a fixed amount of dilation values (relative to the length of the input time series) that are assigned to the kernels in the set. Moreover, MINIROCKET embraces the importance of the PPV operation and relies on it as the sole pooling operation for feature maps.

Applying the MINIROCKET transformation, we use the following procedure to create a high-dimensional representation of the input time series for each modality:

(1) We separately apply MINIROCKET in its default configuration to each univariate input time series, resulting in a representation with 9996 features for each (i.e., the number of PPV features for all kernel-dilation combinations).
(2) Next, we concatenate the resulting MINIROCKET representations created for input time series of each modality of facial behavior (i.e., grouped by Head Position, Head Orientation, Eye Gaze, and Facial Expressions) into matrices with dimensions #modality-features × 9996. This operation is denoted *Concat-TSR* in Figure 1.
(3) We then aggregate these modality-specific matrices of time series representations by calculating their column-wise mean (μ) and variance (σ^2). This creates a single feature vector for each modality with a dimensionality of 19992 features(i.e. $2 \times \#PPV\,features$). This operation is denoted as *Aggretate-TSR* in Figure 1. Intuitively, the resulting aggregate of the

PPV features provides a measure for the overall presence and dispersion of the pattern encoded by a particular kernel-dilation combination across the time series comprising a particular modality.

(4) Finally, we concatenate the aggregated representations for each modality into a joint multimodal representation (*Concat-Multimodal* in Figure 1) for classification.

In total, this creates a multimodal representation of facial behavior with 79968($= 4 \times 9996 \times 2$) features.

4.4 Classification

Finally, we provide the resulting time series representation to either one of two classifiers for training and inference [1], resulting in two different approaches for recollection detection:

4.4.1 RIDGE:. The original ROCKET algorithm and its derivatives rely on an L_2-regularized linear model (i.e., Ridge Regression) for classification [10]. This choice is motivated by the model's fast training time and the capacity to handle high-dimensional input with many redundant features. Moreover, given the intuitive interpretation of PPV features as the presence of the pattern represented by specific kernels, the assumption of a linear relationship between them and a target in the original ROCKET algorithm seems broadly plausible. Building on this notion, in our aggregated time series representation, individual features can be conceived as indicating either the average presence or overall variation of a specific pattern across a modality. Here, a linear relationship between these features and target classes seems similarly plausible. Together, this makes a Ridge classifier a sensible choice. We apply Standardization (i.e., conversion to Z-Scores) to the aggregated time series representation before classification.

4.4.2 PCA99-SVM:. In this approach, we feed our time series representation into a Support Vector Machine with Radial Basis Function (RBF)-Kernel. SVMs have been widely used in similar user-modeling tasks (e.g., affect detection [12]) and are typically data-efficient, making them a suitable choice for our explorations. Moreover, existing work using ROCKET for behavioral modeling has revealed improved performance when using SVMs over the default Ridge Classifier approach [13]. Before classification, we apply Standardization (i.e., conversion to Z-Scores) and Principal Component Analysis (PCA) on our aggregated time series representation, preserving 99% of the input variance. Decreasing the dimensionality of our data in this way reduces the training time the SVM needs substantially.

5 EMPIRICAL INVESTIGATIONS

This section describes a series of machine learning experiments using our predictive modeling approach to explore the feasibility of detecting recollections from facial behavior. Concretely, we investigate two scenarios with different experiment setups: (1) the potential for *Person-independent (PI) detections* detections, i.e., we evaluate whether either of our two approaches can identify and exploit similarities across individuals' facial behavior to make predictions on data from users that they were not trained on. Additionally, we investigate the potential for (2) *Video-independent (VI) detections*, where we evaluate the capacity of our models to detect recollections occurring in potentially known individuals from their behavioral responses to novel videos not included in the training data. The outcome of either experiment holds immediate relevance for user-modeling in applications. For example, the capacity for person-independent detections is highly desirable for systems since it would enable them to estimate users' recollections without requiring additional person-specific training. In essence, they would work for new users already on the first trial (i.e., by avoiding a type of cold-start problem for predictions [24]). However, applications – especially those providing media content –, can often draw on a stable user base regularly interacting with novel content. Video streaming services are a prime example of this. In such a setting, a system might feasibly have access to past instances of any users' behavior as training data. It can then use this to facilitate personalized predictions based on a user's behavior while watching newly available content for which no training data exists yet (i.e., in a video-independent way). In our experiments for these scenarios, we collect a sample with multiple data points of test performance for both our RIDGE and PCA99-SVM approach. The distribution of these samples forms the basis for statistical analyses in which we compare the performance of our approaches (1) against that of a majority classifier as a baseline to verify improvement over chance predictions, and (2) compare our two against each other. Beyond these comparisons, we conduct additional analyses of the performance of both approaches, focusing on their dependency on person-specific recollection biases for accurate predictions and the importance of individual modalities of facial behavior for their performance.

5.1 Experimental Setup

5.1.1 Person-independent (PI) Detections: For training and evaluation of our models in this series, we rely on a *Nested 10-Fold Leave-Persons-Out Cross-Validation procedure (10-Fold LPOCV)*. This procedure creates 10-folds such that no data points from the same *person* are simultaneously available for both training and evaluation of a model. Consequently, this evaluation provides insights into how well predictions based on behavioral patterns generalize to responses from new participants. The scheme is nested because it uses the assigned training data for a fold to conduct an additional inner 5-fold LPOCV for hyperparameter optimization of the classifier. We repeat this 10-Fold LPOCV-procedure 3 times, grouping the data into a different set of folds at each iteration. Together, this process results in a sample of $N = 30$ trained models. We keep all models trained in this setup and their predictions on the test data for further analysis. In particular, we evaluate their predictive performance using the *Balanced Accuracy* (Acc_B) metric to account for a potential imbalance in the distribution of class labels in our data. It is computed by taking the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate): $Acc_B = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$. Scores lie in the interval of $[0, 1]$, with a classifier exploiting only the prevalence of the majority class scoring $Acc_B = .5$. Overall, this procedure results in a sample of $N = 30$ data points for the performance of our approach at PI detections of recollection on the data of unseen participants, thus providing more robust estimates than a single train-test split.

[1] For both classifies and the PCA, we rely on the existing implementations in the python library *Scikit-Learn* [28]

Table 1: Test-performance (Acc_B) our approaches (*RIDGE* and *PCA99SVM* in the person-independent (*PI*) and video-independent (*VI*) settings, including comparison versus Majority Classifier (fixed value of $Acc_B = .5$). Significant improvements are highlighted in bold.

		Acc_B		vs. Majority		
Exp.Setting	*Approach*	$M(SD)$	Min/Max	ΔM	$t(df)$	p
PI	**RIDGE**	**.52 (0.044)**	**.434/.605**	**+.020**	**2.544 (29)**	**.0165***
	PCA99SVM	.512 (0.033)	.451/.563	+.012	2 (29)	.055
VI	**RIDGE**	**.589 (0.028)**	**.541/.642**	**+.089**	**17.14 (29)**	**<.001*****
	PCA99SVM	**.623 (0.037)**	**.526/.705**	**.123**	**18.06 (29)**	**<.001*****

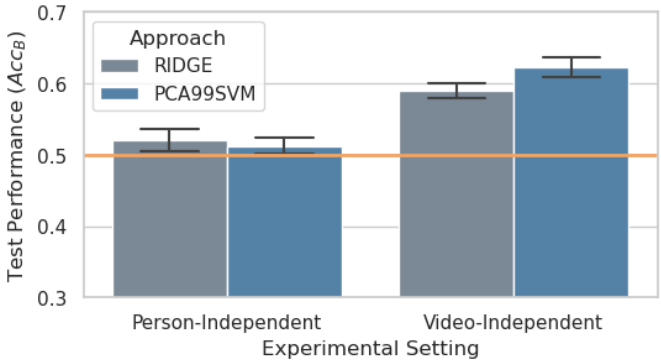

Figure 2: Test-performance (Acc_B) of our *RIDGE* and *PCA99-SVM* approaches in the person-independent (*PI*) and video-independent (*VI*) settings. The orange line demarcates the performance of a Majority Classifier.

5.1.2 Video-independent (VI) Detections: For this series of experiments, we follow a *Nested 10-Fold Leave-Videos-Out Cross-Validation (LVOCV)* procedure, which is repeated 3-times. It includes the same steps as described above, except that it ensures that no responses of participants to the same *video* are present across folds. Note that this procedure results in video-independent but not person-independent detections because of how our dataset has been collected. In particular, responses from the same person are associated with different videos and may be shared across training and test folds. Overall, this procedure results in $N = 30$ trained models and a sample of $N = 30$ measures of their associated test performance for VI detections of recollections (in Acc_B).

5.2 Results and Analysis

In the following, we describe the results for person-independent (PI) and video-independent (IV) detections and additional analysis and experiments related to their performance. We show a graphical overview of the test performance displayed by the two variants of our architecture for both settings in *Figure 2*).

5.2.1 Comparison against Majority Classifier: To verify the feasibility of automatic detections with our approach we compare the displayed performance to that achieved by a majority classifier (i.e., $Acc_B = .5$) using statistical analyses with one-sample t-Tests (see *Table 1.* This comparison indicates a slight improvement over the majority classifier baseline for detections in a PI setting for RIDGE. At the same time, our PCA99-SVM approach appears to display no substantial performance in this setting. In contrast, our approaches perform substantially above the majority baseline at VI detections. Together, these results confirm the feasibility of automatic detections of spontaneous recollections in a VI setting while casting doubt on the ease of detections in the more generic PI setting.

5.2.2 Comparison between PI and VI detections: Statistical comparison using two-sample t-tests between the average performance of our approaches within each experimental setting confirm that PCA99-SVM outperforms RIDGE at video-independent ($\Delta M = .03$, $t(54.187) = 3.976$, $p < .001$), but not at person-independent predictions ($\Delta M = -.008$, $t(53.764) = -0.837$, $p = .407$).

5.2.3 Dependence of VI detections on Person-specific Recollection Bias: One potential explanation for the better performance in VI detections is that models might have learned to identify specific individuals in the training dataset and exploit their overall tendency for recollection. In the following, we describe a series of experiments to explore the dependence of our models on exploiting such person-specific biases in the training set.

After transforming the video data into a time series of behavioral signals, all visual information is lost; only properties consistently present across a person's time series data could feasibly indicate their identity to our models. In particular, we want to rule out the effect of any similarity in participant position influencing signal amplitude or range (e.g., a person always positioned close to the camera might consistently have a greater apparent range of motion). For this purpose, we use two different transformations on univariate time series to generate invariant representations in participant position or movement variance. The one is (1) *Centering* (subtracting the mean amplitude of a time series from that at each time step), and the other (2) *Standardization* (centering and then dividing by

the standard deviation). See the relevant markers in *Figure 1* for the locations in our overall architecture where we integrate these operations.

We collect 30 performance samples using these transformations for both the RIDGE and the PCA99-SVM approach by training them in the VI prediction setting described in *Section 5.1.2* above. We re-use the same cross-validation splits as with the untransformed time series. This process creates data for the following three conditions: the original models without any transformation (*None*), when centering is applied (*Center*), and when standardization is applied (*Standard*).

Paired t-tests between these performance samples reveal that centering does not significantly decrease the performance of RIDGE but does hamper that of PCA99-SVM($\Delta M = -.032, t(29) = -3.568$, $p =< .001$.). In contrast to this, standardization results in a significant decrease in performance for both approaches (RIDGE: $\Delta M = -.045, t(29) = -5.925$, $p =< .001$.; PCA99-SVM: $\Delta M = -.038$, $t(29) = -4.436$, $p =< .001$.). These results quantify the impact of destroying information that models could use to exploit person-specific recollection tendencies in predictions. However, note that these transformations also may eliminate aspects of facial behavior that are likely legitimate for predictions (e.g., whether someone is leaning in to pay closer attention throughout a video). As such, transformations could harm performance beyond what could reasonably be attributed to preventing the exploitation of biases. In the following, we describe additional analyses to investigate this potential impact.

For this purpose, we first compute an indicator for the average imbalance in the training examples for each participant across the folds of the LVOCV scheme (i.e., their person-specific recollection bias as available to the model). It consists of the absolute distance of the observed proportion of positive training samples for a participant from a perfectly balanced proportion of .5. This measure is constrained to the interval $[0, .5]$ and is high for a strongly imbalanced dataset. In addition, we use the predictions of the trained models to compute the per-participant-average accuracy achieved across all folds per distortion condition. Using these measures, we then conduct a statistical analysis of the differences in the person-specific accuracy (*persAcc*) across the three different transformation conditions ($transformType$; factor with 3 levels) while controlling for the average recollection bias of a person in the relevant training data (*persRecBias*) as a covariate. Because data from the same participants are present across the conditions, observations are no longer independent (i.e., repeated measures). For this reason, we rely on linear mixed-effects models, which can account for this nested structure by including random effects (i.e., where supported by the data, we include participant identity as a factor with 295 levels when fitting models). Results reveal a significant effect of the transformations on the average participant-specific accuracy for both PCA99-SVM ($F(2, 590) = 64.995$, $p < .001$, $\Delta R^2 = .093$) and more strongly for RIDGE ($F(2, 885) = 157.108$, $p < .001$, $\Delta R^2 = .258$). Posthoc comparisons of the estimated marginal means reveal that there is no significant difference in the person-wise average accuracy between the Center- and the None-condition (RIDGE: $\Delta M = .0, t(593) = 0$, $p = 1.$); PCA99-SVM: $\Delta M = -.02, t(592) = -2.105, p =< .107$). This finding indicates that centering has no substantial effect on the accuracy of predictions for a person when controlling for imbalances in their training data. In contrast, there is a significantly lower performance in the Standard-condition for both approaches (RIDGE: $\Delta M = -.31, t(593) = -15.317$, $p =< .001$., SVM: $\Delta M = -.149, t(593) = -10.716$, $p =< .001$.). This strong drop indicates standardization harms performance beyond what could be attributed to imabalanced training data.

These findings suggest that centering destroys information about person-specific imbalances more specifically than standardization. Consequently, it seems plausible that the degree to which performance of our approaches depends on exploiting person-specific recollection biases resides between the None- and Center-conditions (i.e., around 0% for RIDGE, and around 3% for PCA99-SVM). Moreover, the results indicate that mere exploitation of person-specific biases in recollection is not the primary driver for the performance of either approach in the VI-setting.

5.2.4 Contribution of Individual Modalities to VI detections: To understand the relative contributions of the different aspects of facial behavior, we conduct systematic ablation studies. We limit these explorations to the VI setting since our approaches displayed low predictive accuracy at PI detections. Concretely, to investigate the informative value of specific modalities, we generate new performance samples for either of our approaches with the same splits of the repeated 10-fold LVOCV scheme used to evaluate their overall performance. However, instead of training and testing models of both approaches with the full input data, we drop all features belonging to one particular modality. We repeat this process for each modality in turn, i.e., we first drop all features for Head Rotation (*w/o R*), then Head Position (*w/o P*), next Eye Gaze (*w/o G*), and finally Facial Expressions (*w/o E*). This procedure results in an additional set of $N = 30$ trained models and performance samples for each modality per approach. Comparing the performance differences to samples from the full model (*R+P+G+E*) provides insights into a modality's unique information, with modalities making more unique contributions to the overall performance resulting in a steeper drop. We show a statistical analysis of the difference in performance in Table 2 and graphical depiction in Figure 3. The results demonstrate that removing most modalities does not significantly change performance. The sole exception are Facial Expressions (E), denying access to which reduces performance (significantly so in the PCA99-SVM approach, but with a similar trend for RIDGE). This finding indicates that facial expressions contribute unique information for the detection and cannot be easily compensated by any other modalities. This finding is a contrast to the remaining modalities. Notably, there is no significant drop in performance when Eye Gaze (G) is removed, suggesting that Head Position (P) and Head Rotation (R) provide largely the same information to gaze behavior in this setting.

6 DISCUSSION

6.1 Empirical Findings

Our empirical investigations explore the possibility of automatically analyzing facial behavior to detect the occurrence of spontaneous recollections. Concretely, our goal is to assess whether it is possible to use facial behavior to differentiate between responses to a video that involve recollections and those that do not. We investigate this

Table 2: Test-performance (Acc_B) of the modality-wise ablations for each of the two variants of our approach (*RIDGE* and *PCA99SVM* in the video-independent (*VI*) experiment setting. Includes statistical comparisons between models without the indicated modality (e.g. *w/o R* is without Head Rotation) and models with access to all modalities. Significant differences are highlighted in bold.

		Acc_B		vs. G+P+R+E		
		$M(SD)$	Min/Max	ΔM	$t(df)$	p
RIDGE	**R+P+G+E**	**.589 (0.028)**	**.541/.642**	–	–	–
	w/o R	.585 (0.030)	.533/.686	-.004	-0.646 (29)	.523
	w/o P	.587 (0.035)	.507/.650	-.002	-0.268 (29)	.791
	w/o G	.592 (0.031)	.528/,644	+.003	0.294 (29)	.771
	w/o E	.571 (0.037)	.505/.656	-.018	-1.873 (29)	.071
PCA99SVM	**R+P+G+E**	**.623 (0.037)**	**.526/.705**	–	–	–
	w/o R	.619 (0.036)	.542/.693	-.004	-0.38769 (29)	.701
	w/o P	.609 (0.037)	.532/.684	-.014	-1.5652 (29)	.128
	w/o G	.606 (0.027)	.564/.655	-.017	-1.9284 (29)	.063
	w/o E	**.593 (0.033)**	**.518/.664**	**-.030**	**-4.409 (29)**	**<.001*****

R: Head Rotation; P: Head Position; G: Eye Gaze; E: Facial Expressions; w/o: trained without modality

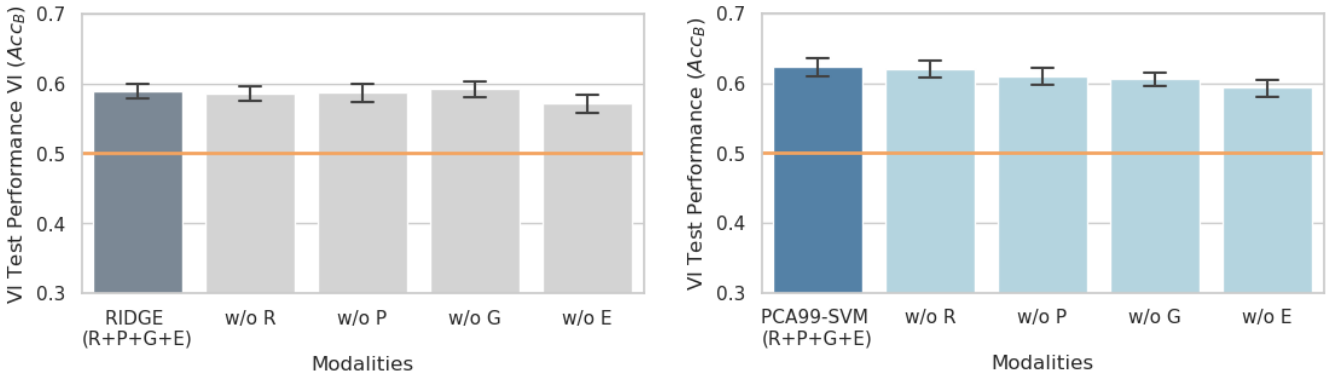

Figure 3: Test-performance (Acc_B) of the modality-wise ablations for each of our two approaches *RIDGE* and *PCA99-SVM*) in the video-independent (*VI*) experiment setting. The orange line demarcates the performance of a Majority Classifier.

possibility in a person-independent (PI) and a video-independent (VI) setting.

Our findings suggest that PI detection may be possible, at least in principle, but is a very challenging task: we observe a very low average performance in this setting, with only one of our approaches performing significantly above the majority baseline. In contrast to this, the results of the VI experiments are promising for user-modeling and personalization research. They indicate the principal possibility of detecting whether a particular video stimulus has triggered personal memories in viewers under challenging in-the-wild conditions. Moreover, their performance suggests the existence of person-specific behavioral patterns that can potentially be learned from relatively few observations of a particular user's responses. Despite this, it is unclear what role the video stimuli play as context and drivers of responses. For example, it might be that people experiencing a recollection fail to respond to certain cues present in the video, which reliably evoke characteristic responses from viewers otherwise. Targeted research is needed to clarify these points. Such a comparison would benefit from a dataset that includes a greater variation in participants and a greater amount of responses from each (Mementos only provides up to 7 unique responses per participant). While our models appear not to substantially exploit invalid person-specific biases when making predictions, future study datasets would be constructed in ways that intentionally safeguard against this possibility. For example, responses from the same participant could be recorded spaced out in time and taking place in different viewing conditions.

Finally, our investigations of the relative importance of different modalities for predictions reveal that facial expressions hold unique insights into the occurrence of recollections. This finding is broadly in alignment with El Haj et al. [17], who identified a connection between facial expressions and episodic memories. However, facial activity is overall sparse in the responses captured by the Mementos dataset (see the relevant descriptions in the official dataset publication [16]). Consequently, it might be a relatively strong indicator for recollections when any facial activity does occur. In stark contrast, eye gaze did not contribute much unique information to predictions in our setting. This finding is surprising, given that much of the related work on internally-directed attention and mind-wandering has repeatedly identified gaze behavior as an indicator [1]. Notably, in our investigation, there seems to be no added benefit of accessing information about a person's eye gaze compared to looking at the overall orientation of their head. A possible

explanation for this could be the challenging setting captured by the dataset, with substantial variation in camera angle, distance to the camera, and lighting conditions. These conditions may result in noisy gaze detections that do not facilitate effective predictions. Moreover, many aspects of eye-related behavior indicative of some forms of internally-directed attention, e.g., pupil dilation, are not accessible using our approach [1]. Nevertheless, the possibility of substituting eye gaze with head orientation features for detecting recollections could benefit deployment in real-world applications since the latter are arguably less sensitive to certain conditions that are prominent in the wild (e.g., occlusions or reflections by glasses). Future research should explore this connection more closely. Beyond application benefits, such investigations might shed light on the role of facial behavior and its different aspects as markers of internally-directed attention more generally. In particular, the majority of computational work on internally-directed attention has focused on stimuli or tasks involving a strong visual component, such as reading [6], driving [38], or film comprehension [22]. This focus could be why eye behavior has been identified as a reliable source for automatic detections (see the relevant discussion of this point by Annerer-Walcher et al. [1]). The comparatively low importance of capturing eye gaze behavior for detections in the music-oriented setting that we investigate might indicate alternative behavioral patterns when other sensory modalities are more important. Datasets capturing the evocativeness of stimuli across different modalities could shed further light on the role of different behavioral components and their overall use for detecting recollections.

6.2 Limitations

Our findings have several limitations rooted in our technological approach and the dataset we relied on for our empirical investigation. Regarding technology, we only provide explorations with a particular computational architecture as a proof-of-concept. While it serves our research goal of exploring the principal feasibility of automatic recollection detection, our comparison offers only limited insights into the overall difficulty of the task, e.g., with different algorithms for time series representation and classification. Notably, large-scale comparisons demonstrate a substantial degree of variability in the performance of any one particular approach for representing time series across tasks, with large-scale ensembles of different approaches consistently performing best (albeit at great computational expense) [2, 11]. We have motivated our decision to choose MINIROCKET based on its performance and overall time and data efficiency. However, other representations might lead to better results, and future work could explore a wider variety of different modeling approaches. Notably, our experiments reveal poor performance in the PI-setting, pointing towards two potential causes: (1) that our specific approach is ill-suited to exploit any existing person-independent behavioral markers effectively, or (2) that patterns of facial activity indicating recollections in a person-independent fashion are sparse and comparatively uninformative. An indication that our approach for this task might limit performance is that related work on detecting mind wandering and internally-directed attention has repeatedly identified potential for person-independent predictions [1, 6]. However, some of this work has been undertaken in more stable laboratory conditions, aims for predictions at a finer-grained temporal resolution, and involves different task settings (e.g., detecting attentional drifts during reading [6] or interacting with a tutoring system [7]). Moreover, there are also existing computational studies that underline the challenge of person-independent predictions of mind-wandering (e.g., based on driver behavior [38]). Overall, this state of affairs makes it difficult to compare existing findings to ours directly and requires further research focusing on recollections. However, the comparatively strong performance of our approach in the VI experiments suggests both the existence of person-specific behavioral markers and the capacity of our approach to pick them up. A straightforward way to shed further light on the dependence of the results on our particular approach would be to evaluate the performance of alternative architectures in both PI and VI experiment settings on the Mementos dataset.

Beyond technological aspects, there might be additional limitations resulting from our dataset. For once, the video segments to which participants in Mementos respond are short (about one minute). This duration might not be enough for spontaneous thoughts with a prolonged or pronounced behavioral component to emerge for automatic analysis to register. Moreover, the recollections captured in the dataset are not coded in terms of their specific phenomenological properties. For example, whether they are primarily episodic (involving mental time traveling back to specific events) or semantic (mainly language-based factual knowledge about one's past). However, these different types have been associated with variations in the degree of facial expressions that they evoke [17]. Additionally, the dataset does not lend itself to studying whether patterns of facial behavior indicative of recollections meaningfully differ from those for other forms of internal attention (such as creative thinking or goal-directed thought [8]). Consequently, future explorations of detecting recollections from facial behavior would benefit from additional datasets capturing both (1) responses to various kinds of media stimuli in a wider variety of situations (e.g., photographs, text, or musical pieces), and (2) self-reports differentiating between various forms of internally directed attention in general and types of recollection in specific. Such corpora would substantially support progress towards detecting the evocativeness of stimuli and situations, and they would also complement existing efforts in modeling states of internally-directed attention more generally. Finally, the scale of the Mementos dataset limits the types of algorithms that can be meaningfully applied to it (e.g., for deep learning). As such, the development of larger corpora capturing recollection processes in response to stimuli would be a valuable asset for further technological research.

7 SUMMARY AND CONCLUSION

The personal memories that we associate with people, places, and even media content are an important ingredient for what makes these things take on significance in our present lives. For this reason, recollections can strongly influence our emotional experience in the presence, which, in turn, can define our responses and attitudes towards whatever triggers them [15]. As such, detecting the occurrence of recollections might facilitate powerful potential capabilities for intelligent systems, such as learning about a person's

unique preferences (e.g., for a particular video) and understanding where these come from (e.g., it is because of personal memories and not due to stylistic preferences).

In this article, we have explored a first step towards leveraging this potential by demonstrating the principal feasibility of automatically detecting whether a video has triggered recollections in a viewer based on their facial behavior. For this purpose, we have developed a predictive modeling architecture that we use in a series of machine learning experiments to investigate feasibility in two different settings that are relevant for applications. One of these is person-independent (PI) detection, where ideally, no data from any specific user would be required for training to analyze their responses accurately. The other setting is video-independent (VI) detection, where accurate estimates of occurring recollections should be possible, at least for responses of an existing user-base to new video stimuli, without the system first needing to be trained on associated data for that stimulus. Our results show performance above chance level in both settings. However, detection performance was overall poor in the PI setting, indicating that detections without personalization might be challenging. However, the performance of our approaches in the VI setting is promising. Since predictions in this setting approximate detections based on an existing user base's responses that are exposed to new video content, this is a finding of direct relevance for application contexts where similar conditions hold (e.g., media content platforms). Overall, our findings indicate that recollections detection based on visual analysis of facial behavior is principally possible, even in technologically challenging in-the-wild conditions. Moreover, in contrast to findings in existing research, analysis of our models for VI detections reveal that their performance does not substantially depend on access to gaze behavior. Instead, they indicate that gaze can be reasonably well substituted with more coarse information about head rotation and position alone. Because of this, automatic recollection detection might be able to side-step some of the challenges faced by gaze estimation in the wild (e.g., reflections, eyewear, or angle and distance to the camera), further adding the robustness needed for real-world applications. Nevertheless, while our findings are promising and indicate a potential for progress towards applications in the wild, they also underline the challenging nature of the task and the need for further data collection and technological research.

ACKNOWLEDGMENTS

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture, and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

REFERENCES

[1] Sonja Annerer-Walcher, Simon M. Ceh, Felix Putze, Marvin Kampen, Christof Körner, and Mathias Benedek. 2021. How Reliably Do Eye Parameters Indicate Internal Versus External Attentional Focus? *Cognitive Science* 45, 4 (2021). https://doi.org/10.1111/cogs.12977

[2] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. 2017. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery* 31, 3 (may 2017), 606–660. https://doi.org/10.1007/s10618-016-0483-9

[3] 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

[4] Krystian Barzykowski and Agnieszka Niedźwieńska. 2018. Involuntary autobiographical memories are relatively more often reported during high cognitive load tasks. *Acta Psychologica* 182 (jan 2018), 119–128. https://doi.org/10.1016/j.actpsy.2017.11.014

[5] Dorthe Berntsen, Søren Risløv Staugaard, and Louise Maria Torp Sørensen. 2013. Why am I remembering this now? Predicting the occurrence of involuntary (spontaneous) episodic memories. *Journal of Experimental Psychology: General* 142, 2 (2013), 426–444. https://doi.org/10.1037/a0029128

[6] Robert Bixler and Sidney D'Mello. 2014. Toward Fully Automated Person-Independent Detection of Mind Wandering. In *User Modeling, Adaptation, and Personalization*, Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben (Eds.). Springer International Publishing, Cham, 37–48. https://doi.org/10.1007/978-3-319-08786-3_4

[7] Nigel Bosch and Sidney K. D'Mello. 2021. Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom. *IEEE Transactions on Affective Computing* 12, 4 (oct 2021), 974–988. https://doi.org/10.1109/TAFFC.2019.2908837

[8] Kalina Christoff, Zachary C. Irving, Kieran C. R. Fox, R. Nathan Spreng, and Jessica R. Andrews-Hanna. 2016. Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience* 17, 11 (nov 2016), 718–731. https://doi.org/10.1038/nrn.2016.113

[9] Dan Cosley, Victoria Schwanda, S Tejaswi Peesapaty, Johnathon Schultz, and Jonathan Baxter. 2009. Experiences with a publicly deployed tool for reminiscing. In *Proc. First Int'l Workshop on Reminiscence Systems*. 31–36.

[10] Angus Dempster, François Petitjean, and Geoffrey I. Webb. 2020. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery* 34, 5 (sep 2020), 1454–1495. https://doi.org/10.1007/s10618-020-00701-z arXiv:1910.13051

[11] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. 2021. MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2021), 248–257. https://doi.org/10.1145/3447548.3467231 arXiv:2012.08791

[12] 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

[13] Bernd Dudzik, Simon Columbus, Tiffany Matej Hrkalovic, Daniel Balliet, and Hayley Hung. 2021. Recognizing Perceived Interdependence in Face-to-Face Negotiations through Multimodal Analysis of Nonverbal Behavior. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 121–130. https://doi.org/10.1145/3462244.3479935

[14] 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

[15] 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

[16] Bernd Dudzik, Hayley Hung, Mark A. Neerincx, and Joost Broekens. 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 (2021), 1–18. https://doi.org/10.1109/TAFFC.2021.3089584

[17] Mohamad El Haj, Pascal Antoine, and Jean Louis Nandrino. 2016. More emotional facial expressions during episodic than during semantic autobiographical retrieval. *Cognitive, Affective, & Behavioral Neuroscience* 16, 2 (apr 2016), 374–381. https://doi.org/10.3758/s13415-015-0397-9

[18] Myrthe Faber and Sidney K. D'Mello. 2018. How the stimulus influences mind wandering in semantically rich task contexts. *Cognitive Research: Principles and Implications* 3, 1 (dec 2018), 35. https://doi.org/10.1186/s41235-018-0129-0

[19] James J. Gross and Robert W. Levenson. 1995. Emotion elicitation using films. *Cognition & Emotion* 9, 1 (1995), 87–108. https://doi.org/10.1080/02699939508408966

[20] Kunal Gupta, Sam W. T. Chan, Yun Suen Pai, Nicholas Strachan, John Su, Alexander Sumich, Suranga Nanayakkara, and Mark Billinghurst. 2022. Total VRecall. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6 (7 2022), 1–21. Issue 2. https://doi.org/10.1145/3534615

[21] Jacek Gwizdka. 2019. Exploring Eye-Tracking Data for Detection of Mind-Wandering on Web Tasks. In *Lecture Notes in Information Systems and Organisation*. Vol. 29. Springer, Cham, 47–55. https://doi.org/10.1007/978-3-030-01087-4_6

[22] Michael Xuelin Huang, Jiajia Li, Grace Ngai, Hong Va Leong, and Andreas Bulling. 2019. Moment-to-Moment Detection of Internal Thought during Video Viewing from Eye Vergence Behavior. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, New York, NY, USA, 2254–2262. https://doi.org/10.1145/3343031.3350573 arXiv:1901.06572

[23] 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

[24] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert Systems with Applications* 41, 4 (mar 2014), 2065–2073. https://doi.org/10.1016/j.eswa.2013.09.005

[25] Mei Yii Lim. 2012. Memory Models for Intelligent Social Companions. In *Studies in Computational Intelligence*. Vol. 396. 241–262. https://doi.org/10.1007/978-3-642-25691-2_10

[26] 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

[27] J. O'Rourke, F. Tobin, S. O'Callaghan, R. Sowman, and D. Collins. 2011. 'YouTube': a useful tool for reminiscence therapy in dementia? *Age and Ageing* 40, 6 (nov 2011), 742–744. https://doi.org/10.1093/ageing/afr100

[28] 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.

[29] Marieke M.M. Peeters, Maaike Harbers, and Mark A Neerincx. 2016. Designing a personal music assistant that enhances the social, cognitive, and affective experiences of people with dementia. *Computers in Human Behavior* 63 (oct 2016), 727–737. https://doi.org/10.1016/j.chb.2016.06.003

[30] Felix Putze, Dennis Küster, Sonja Annerer-Walcher, and Mathias Benedek. 2018. 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, 258–262. https://doi.org/10.1145/3242969.3243000

[31] Felix Putze, Maximilian Scherer, and Tanja Schultz. 2016. Starring into the void? Classifying Internal vs. External Attention from EEG. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, Vol. 23-27-Octo. ACM, New York, NY, USA, 1–4. https://doi.org/10.1145/2971485.2971555

[32] Tali Sharot, Matthew L. Davidson, Meredith M. Carson, and Elizabeth A. Phelps. 2008. Eye Movements Predict Recollective Experience. *PLoS ONE* 3, 8 (aug 2008), e2884. https://doi.org/10.1371/journal.pone.0002884

[33] Mohammad Soleymani and Maja Pantic. 2012. Human-centered implicit tagging: Overview and perspectives. *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3304–3309. https://doi.org/10.1109/ICSMC.2012.6378301

[34] Victoria Schwanda Sosik and Dan Cosley. 2014. Leveraging social media content to support engagement in positive interventions. *The Journal of Positive Psychology* 9 (9 2014), 428–434. Issue 5. https://doi.org/10.1080/17439760.2014.910826

[35] Angela Stewart, Nigel Bosch, Huili Chen, Patrick Donnelly, and Sidney D'Mello. 2017. 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, 359–370. https://doi.org/10.1007/978-3-319-61425-0_30

[36] Doménique van Gennip, Elise van den Hoven, and Panos Markopoulos. 2015. 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, 3443–3452. https://doi.org/10.1145/2702123.2702460

[37] Manila Vannucci, Claudia Pelagatti, Maciej Hanczakowski, Giuliana Mazzoni, and Claudia Rossi Paccani. 2015. Why are we not flooded by involuntary autobiographical memories? Few cues are more effective than many. *Psychological Research* 79, 6 (nov 2015), 1077–1085. https://doi.org/10.1007/s00426-014-0632-y

[38] Yuyu Zhang and Takatsune Kumada. 2018. Automatic detection of mind wandering in a simulated driving task with behavioral measures. *PLOS ONE* 13, 11 (nov 2018), e0207092. https://doi.org/10.1371/journal.pone.0207092

[39] Yue Zhao, Christoph Lofi, and Claudia Hauff. 2017. Scalable mind-wandering detection for MOOCs: A webcam-based approach. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10474 LNCS. Springer, Cham, 330–344. https://doi.org/10.1007/978-3-319-66610-5_24