Towards Engagement Recognition of People with Dementia in Care Settings

Lars Steinert
Cognitive Systems Lab
University of Bremen
Bremen, Germany

Felix Putze
Cognitive Systems Lab
University of Bremen
Bremen, Germany

Dennis Küster
Cognitive Systems Lab
University of Bremen
Bremen, Germany

Tanja Schultz
Cognitive Systems Lab
University of Bremen
Bremen, Germany

ABSTRACT

The number of People with Dementia (PwD) is expected to triple by 2050 - and there is no medical cure in sight. The cornerstones of effective secondary therapy are physical, social, and cognitive activation, which require sustained engagement. Recent results indicate that activation can be successfully supported by technical systems. However, to tailor such technical activation systems to the patients' needs, a continuous assessment of engagement is key. Also, the assessment has to be adapted to the patient as activation responses are highly individual. Whereas engagement has been exhaustively investigated in HCI contexts, little is known about the potential of visual cues for the automatic recognition of engagement of PwD. This task is especially challenging because, as faces change with age, they develop wrinkles, folds, and age spots. In this paper, we show that emotional engagement can be automatically recognized based on visual and contextual information derived from interactions with a tablet-based activation system. For this purpose, we took 41 sessions recorded using the I-CARE activation system in an unconstrained care setting, and trained several engagement recognition systems using machine learning techniques. Our evaluation results show that visual and contextual features can successfully be utilized for the recognition of engagement of PwD. We discuss how robust engagement recognition for PwD could contribute to more effective activation systems, and what challenges need to be overcome to ensure long-term acceptance and usability.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Engagement; Dementia; Facial Expressions; Emotion; LSTM

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

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418856

ACM Reference Format:
Lars Steinert, Felix Putze, Dennis Küster, and Tanja Schultz. 2020. Towards Engagement Recognition of People with Dementia in Care Settings. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3382507.3418856

1 INTRODUCTION

Roughly 50 million people worldwide are currently suffering from dementia. This number is expected to triple by 2050 [50]. Dementia is characterized by a loss of cognitive function and changes in behavior. This includes memory, language skills, and the ability to focus and pay attention [50]. However, it has been shown that secondary therapy such as the physical, social and cognitive activation of People with Dementia (PwD) has significant positive effects. Activation impacts cognitive functioning [46, 51] and can help prevent the magnification of apathy, boredom, depression, and loneliness associated with dementia [9]. Furthermore, activation can lead to higher perceived quality of life [10, 43, 45]. We follow Cohen's [9] argument that activation stimuli have to produce engagement to take effect and adopt his definition of engagement as "the act of being occupied or involved with an external stimulus".

As engagement fluctuates over time and is dependent on multiple factors [9], it is crucial for the design of a technical activation system to continuously assess engagement during activation and avoid disengagement by prompting for feedback. Moreover, in later stages of the disease, PwD may no longer be able to explicitly express their needs and feelings. Thus, a non-intrusive continuous assessment of engagement of PwD can help to decide whether to continue, pause, or end an activation. Also, an recommendation system could learn which contents are particularly engaging for the individual PwD and encourage their use. Emotional engagement is one of the most important dimensions of engagement [9, 16]. Therefore, the aim of this paper is to automatically recognize emotional engagement of PwD who used a technical activation system over a period of several months in (unconstrained) care settings. For this, (1) visual features, namely facial expressions, gaze, head pose and, Convolutional Neural Network (CNN) based features, as well as (2) contextual features, namely the self-reported wellbeing of the PwD, the activation content type and whether it was personalized, and the daytime of the activation are used. Based on this, Long Short-Term Memory (LSTM) networks are applied to train

several emotional engagement recognition systems. This paper describes the design, training, and evaluation of the resulting emotional engagement recognition systems for PwD. Moreover, model uncertainty is investigated and compared to human disagreement.

2 RELATED WORK

Numerous studies have examined engagement of PwD in different contexts. Cohen-Mansfield [8] conducted a study on 102 PwD to analyze engagement in activities such as games, storytelling, or choral singing in a group setting, demonstrating that these activities elicited significantly higher engagement than unstructured time, regardless of cognitive functioning. Eggert et al. [12] measured the effects of music and nature images on engagement of PwD, and reported increased engagement in response to both types of interventions. Moyle et al. [28] compared the therapeutic robot seal PARO against a look-alike plush toy to test the effects on engagement of PwD. The study results show that participants using PARO were more verbally and visually engaged than participants who used the plush toy. Also, PARO was more effective in reducing agitation than usual care. Overall, Perugia et al. [35] identified two different methodical approaches towards measuring engagement for PwD: (1) observational rating scales for on-site observations and (2) ethograms and coding schemata which allow for the retrospective annotation based on videos. Multiple works to date have employed one or both of these methods [9, 16, 17, 23, 37, 49].

In the past years, an increasing number of studies have focused on computer-aided interventions, e.g. activation systems or applications for PwD. Olsen et al. [31] introduced "Media Memory Lane", an application that provides nostalgic music and videos for people with Alzheimer's disease. An evaluation of 15 day care clients showed positive effects on engagement, affect, and reduced fidgeting. McAllister et al. [27] designed an application based on the presentation of personalized prompts to elicit engagement. In a first pilot study with three participants, [27] suggested that especially music, old photos, and movie clips could be highly engaging. In a similar vein, the CIRCA system [4] aimed to enhance the interaction between PwD and caregivers by providing video, photographic, or music stimuli via a touch screen system.

Another line of work has furthermore begun to investigate the potential of biosignals for the assessment of user states to improve (activation) applications for PwD and elderly. Alarcao [2] utilized physiological responses such as electrodermal activity or electroencephalography (EEG) to sense the emotional state of PwD and, accordingly, to adapt the selection of displayed contents. A growing number of studies have further suggested facial expressions as a promising indicator for affect of PwD in early and intermediate stages [26, 40], and partially in late stages of the disease [3, 39]. Parekh et al. [33] used video-based information to determine engagement levels of PwD. The authors reported significant correlations with expert rating for some participants. Recently, Ma et al. [25] introduced ElderReact, a multimodal dataset focusing on emotional responses of healthy elders elicited through different stimuli. Their dataset consists of 1,323 short video clips covering six basic emotions. Although these emotions are spontaneous, the authors state that they might be exaggerated due to the setting of data collection which was for a YouTube reaction channel.

Here, we aim for the recognition of spontaneous emotional engagement of PwD in an unconstrained care setting. Whilst the automatic recognition of engagement [29, 47, 52] and emotions [6, 13, 20, 21, 24] has been explored exhaustively in other HCI contexts for healthy individuals, there have only been a few studies so far that have investigated the potential of visual features for the automatic recognition of emotional engagement of PwD. This challenge is not trivial as elderly PwD might differ systematically from typical convenience samples in both the way and clarity they express engagement and emotions [38], as well as in their ability to provide an unambiguous ground truth in the form of subjective verbal responses to different activation types. Simultaneously, this approach is also very promising as the prior work suggests that elderly PwD may substantially benefit from engaging activations - and this might be revealed by a multi-method and multimodal approach towards engagement detection.

3 I-CARE PROJECT

3.1 Project Description

The data used in this study was obtained in the context of the I-CARE project [44]. The I-CARE system uses a tablet to provide user-specific activation contents and is designed to be jointly used by PwD and caregivers in tandems. More specifically, it aims to address individual activation needs and capabilities of PwD. Contents are image galleries, videos, music, phrases, games, texts, or quizzes. Caregivers can furthermore contribute their own personal contents to place greater emphasis on biographical work. The system is mobile and can be used at any location with an internet connection. Recording sessions could therefore take place in the familiar environment of the PwD. Participants for the I-CARE study were acquired in different care facilities in southern Germany. There was no financial compensation for participants of the study, and all participants provided written consent. For inclusion in the study, one of the tandem partners had to fulfill the clinical criteria for a cognitive disorder involving dementia according to the ICD-10 (Alzheimer dementia, vascular dementia, frontotemporal dementia, Korsakoff's syndrome, or Dementia Not Otherwise Specified), ranging from mild to severe. Tandems agreed to take part in at least eight sessions, consisting of one or more activations. In these sessions, PwD were encouraged to wear the Empatica E4 wristband for measurement of physiological signals, while audio and video data were recorded. Fig. 1 shows two activation sessions from the perspective of an external observer.

3.2 Experimental Setup

Activation sessions generally took place in private rooms or in common areas in one of the care facilities. The participants were individual PwD who were accompanied by one or two formal or informal caregivers, e.g. a relative, a nurse, or a volunteer, to form a tandem. As the interactive character of the system presented a novel attraction, activation sessions sometimes drew the attention of other residents. In some cases, this led to spontaneous group sessions in which the camera remained centered on the participant with dementia. While these sessions represented a somewhat more challenging recording situation, they were still included in the

analysis because they represented a natural (albeit rare) response to the introduction of the system.

The tablet (Google Pixel C, 10.2-inch display or Huawei MediaPad M5, 10.8-inch display) was placed on a stand in front of the PwD so that, as far as possible, all faces were aligned with the field of view of the tablet camera. At the beginning of the session, the system asked about the daily wellbeing ("How are you today?") of the PwD using an easy-to-understand smiley rating scale (positive, neutral, negative). After this warm-up, the recommendation system suggested four different activation items, based on biographic or cohort information, personal interest, and previous ratings of the PwD, if available. The system also provided an opportunity to search for specific contents. Next, the tandem partners picked one activation item. After each activation, the system asked the PwD for a rating of how they liked the activation ("Did you enjoy the content?"), again, on a smiley rating scale. Following the rating, the system went directly back to the overview with more recommended activation contents. At this point, the tandem could decide whether or not to continue with another activation.

Figure 1: Two participants and one instructor, who explains the procedure (left). Two participants during an activation session (right). (©AWO Karlsruhe)

3.3 Data Collection

The I-CARE system supports multimodal data collection (video, audio, physiological responses). Video and audio data were collected with the tablet camera and microphone, respectively. Physiological responses were collected using the E4 wristband. Besides these biosignals, the system also collected certain pieces of contextual information such as the activation content type, whether it was personalized, the wellbeing of the PwD, and the daytime the activation session took place. One session corresponded to a sequence of activations. Activation contents usually varied throughout these sessions. In total, 512 sessions with 2,348 single activations were carried out with the I-CARE system from more than 30 participants with PwD. As activation sessions were part of a field study, and we opted for a setup with minimal supervision and setup requirements, several practical obstacles impeded the continuous collection of the full scope of sensor data. Participants were allowed to use the tablet outside of the scheduled study sessions. In these cases, no audio, video, or physiological data were collected for data protection reasons. From the 512 activation sessions recorded in total, 40 % contain both audio and video data. To date, 41 of these sessions have been manually annotated as explained in section 4.1. In this paper, we used these 41 sessions as our dataset. The information about this dataset, relevant to this study, is summarized in Tab. 1. As emotional engagement of PwD is defined via facial emotional responses [16], we focused on visual and contextual features.

Measure	*Value*
Participants (No.)	22
Gender (Female:Male)	12:9
Age (Years)	80.78 (*SD*=8.52)
Total Sessions (No.)	41
Total Activations (No.)	174
Total Frames (No.)	1,444,383
Duration per Session (Min.)	24.23 (*SD*=9.65)

Table 1: Dataset Information.

4 METHODOLOGY

4.1 Engagement Annotation

Engagement was annotated using the "Video Coding - Incorporating Observed Emotion" (VC-IOE) protocol [16]. The VC-IOE covers six dimensions of engagement: emotional, visual, verbal, behavioral, collective, and signs of agitation. Here, we focused on the emotional engagement component which builds on the Apparent Affect Rating Scale (AARS) [38]. Accordingly, emotional engagement is conceptualized on the basis of six discrete categories of affect that can primarily be assessed via facial responses [16]. However, in the I-CARE context, certain negative emotional responses such as anger or fear were not expected. Also, the collected data indicated that the classes neutral and happiness strongly outweighed the other classes. For this reason, we created one class which covers all negative emotions, namely anger, anxiety or fear, and sadness. Ultimately, three classes of emotional engagement were considered, namely positive, neutral and negative. Engagement was annotated frame-wise and retrospectively by two independent raters based on audio-visual data. We computed Cohen's Kappa (κ) between both raters after intensive training on six random test sessions to evaluate inter-rater reliability, and observed a high agreement of κ=0.803 [48]. Tab. 2 shows how often the different emotional responses occur in the dataset. Each of the 41 sessions contains neutral responses displayed by all of the 22 participants with dementia, totalling more than one million video frames. While almost all (37) sessions contain positive responses, only 13 sessions contain negative ones. The latter making up a proportion of only 0.41 % of all recorded video frames.

4.2 Visual Features

The face is arguably the most important non-verbal source for information about another person's affective states [18]. Several studies have demonstrated the usefulness of facial expression measures for affect sensing in HCI [53]. While these work relatively well for posed facial expressions, affect sensing is very challenging for spontaneous facial expressions in the wild [11]. Furthermore, old faces have more wrinkles and folds [14], whilst facial expressions

Emotion	*Sessions (%)*	*Participants (%)*	*Frames (%)*
Neutral	41 (100.0 %)	22 (100.0 %)	1,114,287 (89.68 %)
Positive	37 (90.24 %)	20 (90.91 %)	123,075 (9.91 %)
Negative	13 (31.70 %)	7 (31.82 %)	5,153 (0.41 %)

Table 2: Emotional Engagement (neutral, positive, negative), expressed as frequencies and proportions (%).

become sparse and unclear in advanced stages of dementia [3]. Also, to date there are only a few annotated datasets including elders [25]. Together, these factors pose additional challenges for automatic affect recognition in PwD. In this study, videos were collected at a resolution of 640x480 pixels, with a frame rate of ~30 fps. First, faces were detected, aligned, and cropped from the video recordings using OpenFace 2.0 [5]. As the tablet camera was centered on the participant with dementia, we assumed their face to be visible most of the time. This was also manually checked. Frames without a visible face were excluded from further analyses.

4.2.1 OpenFace Features: OpenFace 2.0 is an open source facial behavior analysis toolkit [5]. It allows for facial landmark detection, head pose and eye-gaze and estimation, and facial action unit (AU) recognition. OpenFace features have successfully been used for related engagement recognition tasks [29, 47, 52]. From the pre-processed dataset, facial features, namely the presence of 18 and the intensity of 17 AUs[1], the location and rotation of the head (head pose), and the direction of eye-gaze were extracted for all frames. Concatenating all mentioned OpenFace features resulted in a 50-dimensional feature vector for each frame.

4.2.2 CNN Features: Based on the successful application of Convolutional Neural Network (CNN) features in recent EmotiW challenges [13, 20, 21, 34], we extracted CNN features using the pre-trained VGG-Face network. VGG-Face is a 16-layer CNN architecture that has been pre-trained for face recognition. All frames were rescaled to 224x224 pixels to match the input size of the CNN, and normalized by subtracting the mean of the dataset. Next, the network was fine-tuned for five epochs using the FER2013 dataset with stochastic gradient descent, a learning rate of 0.0001, and a momentum of 0.9 [13, 20, 21, 24]. Using the output of *fc6* layer resulted in a 4096-dimensional feature vector for each frame. A principal component analysis (PCA) was applied for decorrelation and feature space compression. This lead to a 716-dimensional feature vector explaining 90 % of the variance.

4.2.3 Normalization: Patterns and expressions of facial responses are known to be both highly person-specific, as well as dependent on contexts [42]. We applied an overall z-normalization on all metric and ordinal scaled features to take these interpersonal and intrapersonal variations into account.

[1] AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45. For AU28, OpenFace only provides information about whether the AU is present.

4.3 Contextual Features

Engagement with external stimuli is known to be context-dependent and affected by (1) environmental, (2) person-related, and (3) stimulus-related attributes [9]. Furthermore, contextual factors are likely to be particularly important for the interpretation of responses by PwD, who may be restricted in their ability to express their needs and feelings. Therefore, we considered information available from the 41 activation sessions as context features. This includes:

Daytime (1): Whether the session took place in the morning or afternoon.

Wellbeing (2): Whether the PwD indicated positive, neutral or negative wellbeing at the beginning of the session.

Content Type (3): Whether the content was drawn from an image gallery, video, music, phrase, game, text, or quiz.

Content Personalization (3): Whether the content was personalized or not (e.g. photos and video clips of the family).

Contextual features were frame-aligned and one-hot encoded due to their categorical nature. Concatenating all contextual features resulted in a 14-dimensional feature vector per frame. While daytime and wellbeing remained constant throughout a given session, the activation content types and personalization could change within sessions.

4.4 Engagement Recognition

In general, facial expressions, head pose, and eye-gaze change continuously. LSTMs are a specific form of recurrent neural networks that allow for sequential modeling and, thus the preservation of temporal information. We implemented an LSTM model based on prior work [19, 24, 29, 47, 52]. We used *n-1* subsequent frames for training the LSTM to classify the (annotated) emotional engagement state of an individual target frame *n*. For *n*, we considered a relatively small range (2-30 frames or 0.07-1 seconds) which is in line with the onset of spontaneous smiles [41]. The LSTM consisted of 8 units followed by a softmax layer and a dropout rate of 0.2 for regularization. The model was trained for 50 epochs with an initial learning rate of 0.0001 using Adam Optimization. Models were trained based on pooled data from all participants (1-4 sessions for each participant). Due to the small number of negative responses (0.41 % of all frames), it seemed reasonable to investigate a two-class (Neutral and Positive) and a three-class (Neutral, Positive and Negative) classification approach. For each approach, (1) OpenFace and contextual features (*OCt*) and (2) OpenFace, contextual, and CNN features (*OCtCNN*) were used in an early fusion strategy which promised to work best [25]. Thus, we trained four different classification models. Evaluation was performed session-independently through 10-fold cross-validation on session level. A permutation test was applied for the baseline (*B*) [30]. Model performance was compared to the results from ten random permutations of the training labels. The permutation score provides evidence on whether the results from the classifier are based on random chance, or if an actual connection between data and labels could be found. Micro-averaged F1-score was considered as evaluation metric to assess model performance because it aggregates the contributions of all classes to compute the average score. This metric is especially useful when data is imbalanced [32].

5 RESULTS

5.1 Model comparison

Tab. 3 summarizes the results of the two-class and three-class emotional engagement recognition for both feature sets. For better understanding, t(s) corresponds to the conversion of n to seconds. All models outperform the baselines. This suggests that all models indeed found a connection between features and labels. In the case of the two-class classification, highest scores were reached considering n=5 frames (0.15 s) with $OCtCNN_{2C}$. An F1-Score of 0.83 (SD=0.16) can be reported. The relatively high standard deviation indicates that there are noticeable performance differences between sessions. A closer look at individual sessions reveals that performance for participants who only contributed one session (8 participants) to the dataset is not worse (F1-Score=0.87, SD=0.09), than for participants who contributed multiple sessions (F1-Score=0.83, SD=0.17). This indicates the model's ability to generalize not only to unseen sessions, but also to unseen participants. For the three-class classification, highest scores were reached when classification is based on n=15 frames (0.5 s) with an F1-Score=0.76 (SD=0.19).

n	$t(s)$	OCt_{2C}	$OCtCNN_{2C}$	OCt_{3C}	$OCtCNN_{3C}$
2	0.07	0.80 (0.18)	0.83 (0.17)	**0.72 (0.20)**	0.74 (0.18)
5	0.15	0.81 (0.18)	**0.83 (0.16)**	0.72 (0.29)	0.72 (0.21)
10	0.30	0.79 (0.17)	0.82 (0.19)	0.72 (0.23)	0.70 (0.23)
15	0.50	**0.82 (0.18)**	0.81 (0.19)	0.71 (0.22)	**0.76 (0.19)**
30	1.00	0.81 (0.18)	0.82 (0.19)	0.69 (0.21)	0.71 (0.17)
B	-	0.61 (0.13)	0.66 (0.13)	0.55 (0.14	0.60 (0.15)

Table 3: Emotional engagement two- and three-class classification over segment lengths n using LSTMs. The baseline is represented with B. Values represent the mean (standard deviation) of the micro-average F1-Score over all sessions.

Fig 2 shows the confusion matrices of all best models corresponding to Tab. 3: OCt_{2C} (top left), $OCtCNN_{2C}$ (bottom left), OCt_{3C} (top right) and $OCtCNN_{3C}$ (bottom right). As expected, the performance of all three-class classification models is worse (F1-Score=0.72, SD=0.20; F1-Score=0.76, SD=0.19) than for the two-class problem (F1-Score=0.82, SD=0.18; F1-Score=0.83, SD=0.16). When only using OpenFace and contextual features, the model is able to successfully discriminate all three types of responses (Fig 2 (top right)) despite the imbalance in our dataset (see Tab. 2). While considering CNN features increased the overall performance for the three-class classification, it surprisingly leads to many misclassifications of negative as neutral responses. This type of errors was also reported in other studies [21, 24].

However, overall CNN features added value to the performance of both classification approaches. This finding is in line with other studies which reported the benefits of using CNN features [6, 13, 24] and especially when they are based on the fine-tuned VGG-Face network [13, 20, 21]. It is also noteworthy that using too long histories (large n) decreases the classification performance. This seems plausible as our dataset consists of spontaneous emotional responses which typically have an average onset of 0.59 s [41]. Also, a genuine "slow" smile might take about half a second, after which the expression has passed its apex [22]. This corresponds nicely with the performance peaks at n=15 in Tab. 3 for our videos with a frame rate of ~30 fps.

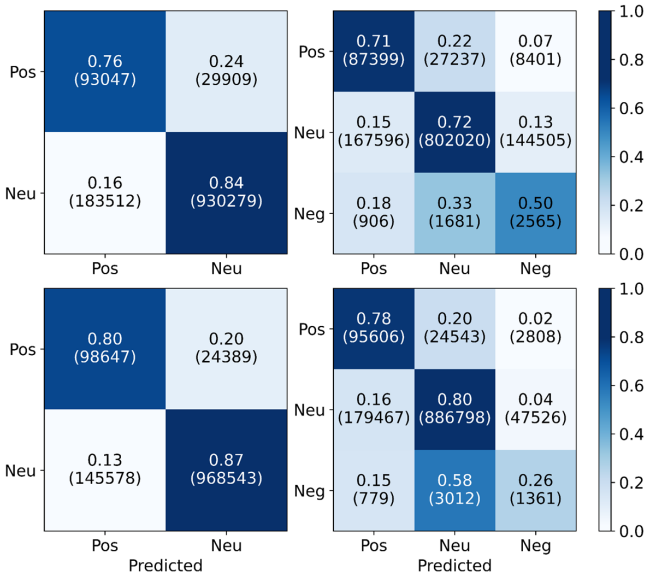

Figure 2: Confusion matrix for OCt_{2C} (top left), $OCtCNN_{2C}$ (bottom left), OCt_{3C} (top right), $OCtCNN_{3C}$ (bottom right). The number of samples is reported in parenthesis.

5.2 Model Uncertainty

Knowledge of the certainty of a classifier can tell a system that is processing the result if it is reliable enough to act upon. In cases of high uncertainty, the system might, for example, decide to collect more samples rather than committing to an unreliable classification. Alternatively, estimates of uncertainty might help to determine the best moment for an intervention, or the timing of an explicit interaction with the user. To estimate (un)certainty, we applied Monte Carlo Dropout to obtain model uncertainty, which is approximated by randomly dropping out hidden units with a certain probability during repeated testing [15]. We used a dropout rate of 0.2 and 50 repetitions. Model uncertainty is reported as the variance in the softmax output. A new model based on a OCt_{2C} with n=15 was trained and tested on all six sessions on which inter-rater reliability (κ=0.803) was calculated. Cohen's κ on these sessions shows a moderate agreement (κ=0.485 with rater 1, κ=0.452 with rater 2) between classifications and both raters [48]. These results are in line with studies on emotion recognition of healthy elders, with a reported κ=0.38 for positive responses [25]. Interestingly, model uncertainty is significantly higher (t=-47.67, p<.001) for frames in which there is disagreement between human raters. This could indicate that there were patterns of facial expressions that were similarly difficult to recognize for both, humans and our model. Also, we observed significantly higher (t=-85.01, p<.001) uncertainty for samples that were misclassified by our model which implies that uncertainty can help to identify likely errors of the classifier. Fig. 3

shows the distribution of model uncertainty for both for inter-rater agreement (left) and classification result (right).

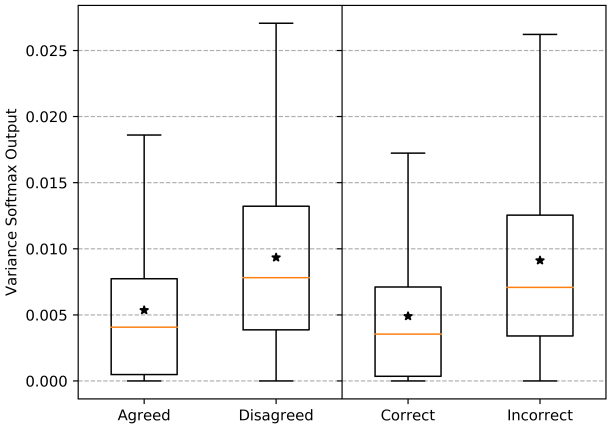

Figure 3: Boxplot with variance of the softmax output for inter-rater agreement (left) and classification result (right). The asterisks represent the statistical mean.

6 DISCUSSION

In the present work, we examined the possibilities of automatic emotional engagement recognition of elderly PwD in an unconstrained care setting, and involving an active role of a tandem partner. This scenario has posed certain unique challenges for our evaluation: PwD often are limited in their capacity to express themselves, and recognition of facial features is obstructed by age-related changes, as well as possibly by dementia-induced changes or comorbid conditions. Furthermore, the presence of the tandem partner in our scenario could be regarded as confounding as they have certainly played a role in the way that participants responded to our activation system. However, we argue that their role in this scenario was, in fact, essential rather than optional. We believe that the social context of the tandem situation is greatly beneficial to PwD as the social interaction is an important factor for their quality of life [7]. Our present results show that facial expressions of PwD in combination with contextual information can be used to automatically recognize emotional engagement of PwD in unconstrained care settings. Also, we obtained our results at a very fine-grained level of individual frames. This provides the opportunity for immediate actions of the activation system in the case of disengagement. We believe that it is important to better understand the dynamics of engagement responses in elderly PwD as based on real-time behavioral data. Our ongoing manual annotation of engagement responses for this dataset makes an important additional contribution by building on the still very sparse body of annotated datasets for elderly PwD.

Nevertheless, there are also certain limitations of our study. Firstly, additional data (audio and physiological signals) should be considered for more robust engagement recognition. In the present work, we decided to focus on the video data, as emotional engagement is primarily assessed via facial expressions [16]. Also, this modality provides the cornerstone of our engagement data. However, emotion recognition from speech is a widely researched area in the fields of spoken communication, HCI and paralinguistics, to name a few [1]. Furthermore, electrodermal activity has shown a potential to provide added value to engagement recognition for PwD [36]. The latter might be especially useful when PwD are no longer able to verbally or facially express their feelings. Future studies could explore the combination of all modalities to compensate for this lack of expression.

While emotional engagement is one of the most important dimensions for engagement [9], further work that examines other aspects of engagement [16] that were beyond the scope of the present study is still required. The recognition of visual engagement, for example, can help to identify the trigger of emotional responses in this unconstrained setting. This, in turn, could shed light on the importance of the tandem partner for success of the activation sessions. Likewise, it would be very interesting to assess engagement in a more predictive manner. A challenging next step in this endeavour could be to not only recognize engagement but to predict engagement in future points in time. Ultimately, an activation system that is able to anticipate the level of engagement would be able to make more sound decisions on whether to continue, pause, or end the activation session. If successful, such a system could then itself improve the impact on engagement of PwD, and thus allow deeper insights into the dynamics of technologically enhanced activation systems that we expect could substantially outperform the present generation of such systems.

7 CONCLUSION

Engagement recognition for elderly PwD is a challenging yet very rewarding endeavour. People with Dementia generally cannot be expected to provide the same level of granularity and frequency of subjective evaluations as typical student convenience samples. Perhaps more importantly from a methodological perspective, the PwD who might most be able to benefit from an engaging activation system may require additional help and social support from a partner. In consequence, our tandem approach does not allow for a clean separation of effects of the system and contributions made by the tandem partner. However, we argue that the notion of designing completely unassisted technical activation systems for people with (advanced) dementia may be misguided. Instead, the presence, contribution, and interest of tandem partners is not a methodological artefact or source of "noise" but rather an integral component of successful activation systems for PwD. Here, we showed how such a tandem-approach may yield meaningful moments of engagement that can be successfully recognized via machine learning techniques despite all challenges. Thus, while dementia is frequently associated with flattened affect, the joint interaction with the tandem partner and the activation contents overall showed very promising results.

ACKNOWLEDGMENTS

This work was partially funded by the Klaus-Tschira-Stiftung. Data collection and development of the I-CARE system was funded

by BMBF-funding body under reference number BMBF-number V4PIDO62.

REFERENCES

[1] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56 – 76. https://doi.org/10.1016/j.specom.2019.12.001

[2] S. M. Alarcão. 2017. Reminiscence therapy improvement using emotional information. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 561–565.

[3] Kenneth Asplund, Astrid Norberg, Rolf Adolfsson, and Howard M. Waxman. 1991. Facial expressions in severely demented patients—a stimulus–response study of four patients with dementia of the Alzheimer type. *International Journal of Geriatric Psychiatry* 6, 8 (1991), 599–606. https://doi.org/10.1002/gps.930060809 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/gps.930060809

[4] Arlene J Astell, Maggie P Ellis, Lauren Bernardi, Norman Alm, Richard Dye, Gary Gowans, and Jim Campbell. 2010. Using a touch screen computer to support relationships between people with dementia and caregivers. *Interacting with Computers* 22, 4 (2010), 267–275.

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

[6] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. 2016. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 433–436.

[7] Hanneke Beerens, Bram de Boer, Sandra Zwakhalen, Frans Tan, Dirk Ruwaard, Jan Hamers, and Hilde Verbeek. 2016. The association between aspects of daily life and quality of life of people with dementia living in long-term care facilities: a momentary assessment study. *International Psychogeriatrics* -1 (04 2016). https://doi.org/10.1017/S1041610216000466

[8] Jiska Cohen-Mansfield. 2018. The impact of group activities and their content on persons with dementia attending them. *Alzheimer's research & therapy* 10, 1 (2018), 37.

[9] Jiska Cohen-Mansfield, Maha Dakheel-Ali, and Marcia S Marx. 2009. Engagement in persons with dementia: the concept and its measurement. *The American journal of geriatric psychiatry* 17, 4 (2009), 299–307.

[10] Jiska Cohen-Mansfield, Marcia S Marx, Khin Thein, and Maha Dakheel-Ali. 2011. The impact of stimuli on affect in persons with dementia. *The Journal of clinical psychiatry* 72, 4 (2011), 480.

[11] Pasquale Dente, Dennis Küster, Lina Skora, and E. Krumhuber. 2017. Measures and metrics for automatic emotion classification via FACET. In *Proceedings of the Conference on the Study of Artificial Intelligence and Simulation of Behaviour (AISB)*. 160–163.

[12] Julia Eggert, Cheryl J Dye, Ellen Vincent, Veronica Parker, Shaundra B Daily, Hiep Pham, Alison Turner Watson, Hollie Summey, and Tania Roy. 2015. Effects of viewing a preferred nature image and hearing preferred music on engagement, agitation, and mental status in persons with dementia. *SAGE Open Medicine* 3 (2015), 2050312115602579. https://doi.org/10.1177/2050312115602579 arXiv:https://doi.org/10.1177/2050312115602579 PMID: 26770801.

[13] Yingruo Fan, Jacqueline C. K. Lam, and Victor O. K. Li. 2018. Video-Based Emotion Recognition Using Deeply-Supervised Neural Networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 584–588. https://doi.org/10.1145/3242969.3264978

[14] Maxi Freudenberg, Reginald B. Adams, Robert E. Kleck, and Ursula Hess. 2015. Through a glass darkly: facial wrinkles affect our processing of emotion in the elderly. *Frontiers in Psychology* 6 (2015). https://doi.org/10.3389/fpsyg.2015.01476

[15] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 1050–1059.

[16] Cindy Jones, Billy Sung, and Wendy Moyle. 2015. Assessing engagement in people with dementia: a new approach to assessment using video analysis. *Archives of psychiatric nursing* 29, 6 (2015), 377–382.

[17] Cindy Jones, Billy Sung, and Wendy Moyle. 2018. Engagement of a Person with Dementia Scale: Establishing content validity and psychometric properties. *Journal of advanced nursing* 74, 9 (2018), 2227–2240.

[18] Arvid Kappas, Eva Krumhuber, and Dennis Küster. 2013. Facial behavior. In *In: Hall, Judith A.; Knapp, Mark L. (Ed.), Nonverbal communication (S. 131-166). Berlin: de Gruyter, 2013.* de Gruyter, 131–166.

[19] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall. 2018. Prediction and Localization of Student Engagement in the Wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. 1–8. https://doi.org/10.1109/DICTA.2018.8615851

[20] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing* 65 (2017), 66–75.

[21] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko. 2017. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. (2017). arXiv:cs.CV/1711.04598

[22] Eva Krumhuber, Antony S. R. Manstead, and Arvid Kappas. 2007. Temporal Aspects of Facial Displays in Person and Expression Perception: The Effects of Smile Dynamics, Head-tilt, and Gender. *Journal of Nonverbal Behavior* 31, 1 (March 2007), 39–56. https://doi.org/10.1007/s10919-006-0019-x

[23] M Powell Lawton, Kimberly Van Haitsma, and Jennifer Klapper. 1996. Observed affect in nursing home residents with Alzheimer's disease. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 51, 1 (1996), P3–P14.

[24] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. 2018. Multi-Feature Based Emotion Recognition for Video Clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 630–634. https://doi.org/10.1145/3242969.3264989

[25] Kaixin Ma, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M Girard, and Louis-Philippe Morency. 2019. ElderReact: A Multimodal Dataset for Recognizing Emotional Response in Aging Adults. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 349–357. https://doi.org/10.1145/3340555.3353747

[26] Carol Magai, Carl Cohen, David Gomberg, Chris Malatesta, and Clayton Culver. 1996. Emotional expression during mid-to late-stage dementia. *International psychogeriatrics* 8, 3 (1996), 383–395.

[27] Margaret McAllister, Jeanne Dayton, Florin Oprescu, Mary Katsikitis, and Christian M Jones. 0. Memory Keeper: A prototype digital application to improve engagement with people with dementia in long-term care (innovative practice). *Dementia* 0, 0 (0), 1471301217737872. https://doi.org/10.1177/1471301217737872 arXiv:https://doi.org/10.1177/1471301217737872 PMID: 29096546.

[28] Wendy Moyle, Cindy J Jones, Jenny E Murfield, Lukman Thalib, Elizabeth RA Beattie, David KH Shum, Siobhan T O'Dwyer, M Cindy Mervin, and Brian M Draper. 2017. Use of a robotic seal as a therapeutic tool to improve dementia symptoms: a cluster-randomized controlled trial. *Journal of the American Medical Directors Association* 18, 9 (2017), 766–773.

[29] Xuesong Niu, Hu Han, Jiabei Zeng, Xuran Sun, Shiguang Shan, Yan Huang, Songfan Yang, and Xilin Chen. 2018. Automatic Engagement Prediction with GAP Feature. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. ACM, New York, NY, USA, 599–603. https://doi.org/10.1145/3242969.3264982

[30] Markus Ojala and Gemma C Garriga. 2010. Permutation tests for studying classifier performance. *Journal of Machine Learning Research* 11, Jun (2010), 1833–1863.

[31] Richard V Olsen, B Lynn Hutchings, and Ezra Ehrenkrantz. 2000. "Media Memory Lane" interventions in an Alzheimer's day care center. *American Journal of Alzheimer's Disease* 15, 3 (2000), 163–175.

[32] Juri Opitz and Sebastian Burst. 2019. Macro F1 and Macro F1. arXiv:cs.LG/1911.03347

[33] Viral Parekh, Pin Foong, Shengdong Zhao, and Ramanathan Subramanian. 2018. AVEID: Automatic Video System for Measuring Engagement In Dementia. 409–413. https://doi.org/10.1145/3172944.3173010

[34] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. , Article 41 (September 2015), 12 pages. https://doi.org/10.5244/C.29.41

[35] G. Perugia, M. Díaz-Boladeras, A. Català-Mallofré, E. I. Barakova, and M. Rauterberg. 2020. ENGAGE-DEM: A Model of Engagement of People with Dementia. *IEEE Transactions on Affective Computing* (2020), 1–1. https://doi.org/10.1109/TAFFC.2020.2980275

[36] Giulia Perugia, Daniel Rodríguez-Martín, Marta Díaz Boladeras, Andreu Català Mallofré, Emilia Barakova, and Matthias Rauterberg. 2017. Electrodermal activity: explorations in the psychophysiology of engagement with social robots in dementia. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1248–1254.

[37] Giulia Perugia, Roos van Berkel, Marta Díaz-Boladeras, Andreu Català-Mallofré, Matthias Rauterberg, and Emilia Barakova. 2018. Understanding Engagement in Dementia Through Behavior. The Ethographic and Laban-Inspired Coding System of Engagement (ELICSE) and the Evidence-Based Model of Engagement-Related Behavior (EMODEB). *Frontiers in psychology* 9 (2018), 690. https://doi.org/10.3389/fpsyg.2018.00690

[38] M. Powell Lawton, K. Van Haitsma, M. Perkinson, and K. Ruckdeschel. 1999. Observed affect and quality of life in dementia: Further affirmations and problems. *Journal of Mental Health and Aging* 5, 1 (29 May 1999), 69–81.

[39] Hans Ragneskog, Kenneth Asplund, Mona Kihlgren, and Astrid Norberg. 2001. Individualized music played for agitated patients with dementia: analysis of video-recorded sessions. *International journal of nursing practice* 7, 3 (2001), 146–155.

[40] Susanna Re. 2003. Emotionales Ausdrucksverhalten bei schweren demenziellen Erkrankungen. *Zeitschrift für Gerontologie und Geriatrie* 36, 6 (2003), 447–453.

[41] Karen Schmidt, Zara Ambadar, Jeffrey Cohn, and Lawrence Reed. 2006. Movement Differences between Deliberate and Spontaneous Facial Expressions: Zygomaticus Major Action in Smiling. *Journal of nonverbal behavior* 30 (02 2006), 37–52. https://doi.org/10.1007/s10919-005-0003-x

[42] Karen L. Schmidt and Jeffrey F. Cohn. [n.d.]. Human facial expressions as adaptations: Evolutionary questions in facial expression research. 116 ([n. d.]), 3–24. Issue S33. https://doi.org/10.1002/ajpa.20001

[43] Andrea S Schreiner, Eiko Yamamoto, and Hisako Shiotani. 2005. Positive affect among nursing home residents with Alzheimer's dementia: the effect of recreational activity. *Aging & mental health* 9, 2 (2005), 129–134.

[44] Tanja Schultz, Felix Putze, Timo Schulze, Lars Steinert, Ralf Mikut, Wolfgang Doneit, Andreas Kruse, Anamaria Depner, Ingo Franz, Marc Engels, Philipp Gaerte, Sebastian Jünger, Rene Linden, Christof Ziegler, Michael Ricken, Todor Dimitrov, Joachim Herzig, Irene Maucher, Keni Bernardin, and Clarissa Simon. 2018. I-CARE - Ein Mensch-Technik Interaktionssystem zur Individuellen Aktivierung von Menschen mit Demenz.

[45] Gary W Small, Peter V Rabins, Patricia P Barry, Neil S Buckholtz, Steven T DeKosky, Steven H Ferris, Sanford I Finkel, Lisa P Gwyther, Zaven S Khachaturian, Barry D Lebowitz, et al. 1997. Diagnosis and treatment of Alzheimer disease and related disorders: consensus statement of the American Association for Geriatric Psychiatry, the Alzheimer's Association, and the American Geriatrics Society. *Jama* 278, 16 (1997), 1363–1371.

[46] Aimee Spector, Lene Thorgrimsen, BOB Woods, Lindsay Royan, Steve Davies, Margaret Butterworth, and Martin Orrell. 2003. Efficacy of an evidence-based cognitive stimulation therapy programme for people with dementia: randomised controlled trial. *The British Journal of Psychiatry* 183, 3 (2003), 248–254.

[47] Chinchu Thomas, Nitin Nair, and Dinesh Babu Jayagopi. 2018. Predicting Engagement Intensity in the Wild Using Temporal Convolutional Network. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. ACM, New York, NY, USA, 604–610. https://doi.org/10.1145/3242969.3264984

[48] Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med* 37, 5 (2005), 360–363.

[49] K. Wada, Yousuke Ikeda, Kaoru Inoue, and Reona Uehara. 2010. Development and preliminary evaluation of a caregiver's manual for robot therapy using the therapeutic seal robot Paro. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 533–538. https://doi.org/10.1109/ROMAN.2010.5598615

[50] WHO. 2017. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia. [Online; accessed 27-March-2019].

[51] Bob Woods, Elisa Aguirre, Aimee E Spector, and Martin Orrell. 2012. Cognitive stimulation to improve cognitive functioning in people with dementia. *Cochrane Database of Systematic Reviews* 2 (2012).

[52] Jianfei Yang, Kai Wang, Xiaojiang Peng, and Yu Qiao. 2018. Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. ACM, New York, NY, USA, 594–598. https://doi.org/10.1145/3242969.3264981

[53] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 1 (2009), 39–58.