AVEID: Automatic Video System for Measuring Engagement In Dementia
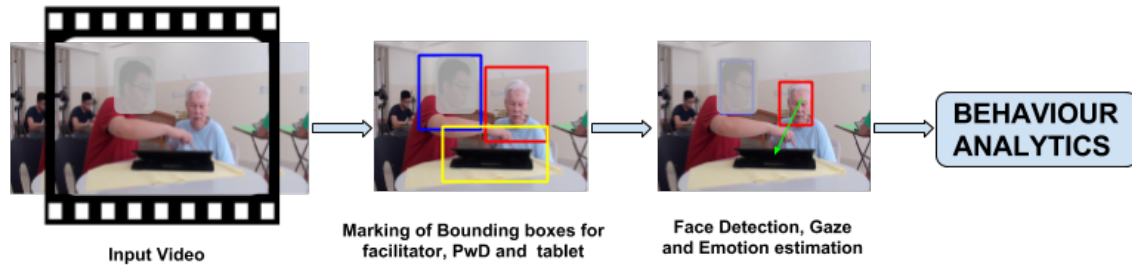

Figure 1 AVEID overview - The AVEID system enables non-intrusive and automated behavioral analytics of persons with dementia (PwD) by capturing their attention (gazing behavior) and attitude (emotion) characteristics.

ABSTRACT

Engagement in dementia is typically measured using behavior observational scales (BOS) that are tedious and involve intensive manual labor to annotate, and are therefore not easily scalable. We present AVEID, a low-cost and easy to use video-based engagement measurement tool to determine the level of engagement of a person with dementia (PwD) when interacting with a target object. We show that the objective behavioral measures computed via AVEID correlate well with subjective expert impressions for the popular MPES and OME BOS, confirming its viability and effectiveness. Moreover, AVEID measures can be obtained for a variety of engagement designs, thereby facilitating large-scale studies with PwD populations.

Author Keywords

Engagement; Video-based analytics; Dementia

ACM Classification Keywords

H.1.2 [User/Machine Systems]: Human information processing; I.5.4 [Pattern Recognition Applications]: Computer vision.

INTRODUCTION

Engagement activities for people with dementia (PwD) are an important, non-pharmacological method to promote quality of life and reduce undesirable outcomes such as apathy, depression and aggressive behaviors [3]. Hence, HCI researchers have been developing various systems to supply more interactive and interesting engagement

activities for PwD. Examples of these are conversation support systems [1], art therapy [11] and music therapy [5].

Equally important are *engagement measurement tools* as they provide feedback to facilitators on the effectiveness of engagement systems, while also providing a basis for forming and adjusting interventions. As memory impairments in PwD preclude the use of self-reports as a measurement tool, researchers primarily use some form of observation to code outcome behavior. Behavioral coding requires the use of behavioral observational scales (BOS), and the training of coders who can accurately encode observed behaviors for robust inference. Due to these requirements, behavioral coding as a measurement tool presents the following challenges: 1) It is human-effort intensive; 2) Training coders is time-consuming; 3) Large-scale data annotation becomes tedious, and 4) It supports only coarse-grained behavior analytics due to limitations in human annotation capability.

Hence, automated measures of engagement might be useful for researchers. Despite the availability of such tools for neurotypical target groups, they are not appropriate for use with PwD. For example, PwD tend to resist any type of on-body physical instrumentation [10] so the use of wearable devices or bio-signal systems for measuring engagement is typically not viable. Patel *et. al.* [14] suggest that PwD require monitoring systems that are "totally unobtrusive, and preferably collected in a transparent way without patient intervention due to their cognitive impairment."

In this regard, we present AVEID, a low-cost and easy to use video-based system for measuring engagement in PwD. AVEID employs deep learning-based computer vision algorithms to unobtrusively and continuously capture a dementia patient's engagement behavior during an interaction session, thereby enabling fine-grained behavior analytics. Consistent with BOS that quantify the patient's *attention* and *attitude* towards an engagement system, AVEID estimates the patient's attentional behavior based on *gazing direction*, and attitude based on *facial emotions* (Figs. 1,6). Also, since deep learning systems are 'end-to-

end' requiring no manual intervention for model synthesis, AVEID only requires manual annotation of bounding boxes to denote positions of the patient, facilitator (if present) and engagement device at the beginning of the examined video. Unlike gaze-tracking or wearable systems that involve specialized hardware, AVEID only requires a video recording as input. These features facilitate practical, day-to-day usage of AVEID in care homes by a variety of users such as therapists or researchers from other domains.

We validated AVEID against human (expert behavioral coder) impressions corresponding to two well-known BOS, namely, the Menorah Park Engagement Scale (MPES) [2] and the Observational Measure of Engagement (OME) [3]. Experiments confirm that measures derived from AVEID agree well with human opinion. Furthermore, AVEID can save the time and effort expended by the behavioral coder, allow for personalized treatment and enable timely analytics on large sample sizes. AVEID measures would also be applicable across small-space engagement activities, facilitating replicability and ecological validity of engagement evaluation studies; we ultimately envision AVEID to provide a strong basis for effective non-pharmacological intervention in dementia care environments.

AVEID IMPLEMENTATION

Consistent with popular BOS used for measuring engagement with PwD, AVEID is designed to quantify the *attention* and *attitude* of the patient towards the engagement system over the observed period (Table 1). AVEID employs the patient's *gaze focus* as a cue towards inferring attention, while utilizing *facial affect* to infer attitude. We use the following terms to describe the system:

User: who operates AVEID to measure engagement.
Target subject: PwD undertaking an engagement activity.
Target activity space: the 2D area where we expect the PwD's gaze to be directed in order to engage with the designed activity. In the AVEID context, the engagement activity involves interaction with a tablet as in [7].
Facilitator: a second person in the frame whose role is to support and promote engagement of the PwD.

AVEID Modules

The AVEID system comprises four modules, namely, *User input*, *Face detection*, *Gaze and emotion detection* and *Behavior analytics* (Fig. 1).

The **User Input** module allows users to select the video for analysis, and enables them to mark bounding boxes corresponding to the target subject, target activity space and facilitator. This initialization needs to be performed at the beginning of each video for accurate face detection (under varied video acquisition conditions), and estimation of attentional measures.

The **Face Detection** module implements the *tiny face* [8] state-of-the-art face detection method. *Tiny face* performs robust face detection across a wide range of illuminations,

face sizes, head poses and facial occlusions. The face detection module detects patient's and facilitator's faces (within the input bounding boxes) in each video frame.

Gaze Detection forms the core of AVEID, as its output is used to compute *attention* measures, which are of prime importance in engagement measurement. This module implements the *GazeFollow* deep network architecture of Recasens *et al.* [15], and utilizes head orientation as a cue to determine where a target is gazing at [13,16].
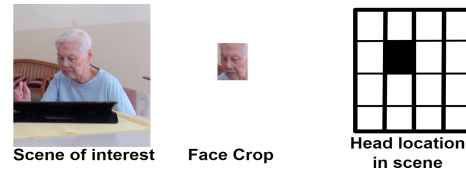

Figure 2. The three inputs utilized for gaze detection.

The target's gaze focus is determined based on three inputs (Fig. 2): 1) An image (video frame) capturing the scene of interest, 2) Cropped head of the target (output of *Tiny face*), and 3) Location of the head in the scene (denoted by the highlighted grid square). The model comprises two computational pathways. The *gaze pathway* uses the target head appearance and location to produce a *gaze map* that estimates the general direction of the target's gaze. The *saliency pathway* examines scene content to output the *saliency map* that detects interesting objects capable of capturing the target's attention. The two maps are then combined to infer the target's gaze focus. As the target activity space and facilitator are the two entities of interest in AVEID, the gaze detection module outputs for each video frame, a label signifying whether the target is gazing at the *target activity space*, *facilitator* or *elsewhere*.

The **Emotion Detection** module implements the deep network for emotion recognition described in [12]. Given that recognizing emotions of elderly people is challenging even for trained human experts, the deep network described in [12] is fine-tuned with 950 elderly face examples from the FACES dataset [4]. The emotion detection module outputs per video frame a label corresponding to one of the six Ekman emotions plus neutral as illustrated in Fig. 6.

The **Behavior Analytics module** processes the outputs of the gaze and emotion detection modules to compute measures reflecting the patient's attention and attitude. For characterizing *attention*, the per-frame gaze labels are combined to compute three raw (or basic) statistics, namely, gaze proportion on tablet, facilitator and elsewhere over the period of observation. In addition to these coarse-grained features, we also derived 18 fine-grained statistics from the gaze labels for analysis, described as follows.

Upon determining *episodes of focus* on the tablet, facilitator and other *entities* within the observation period, we computed the means and standard deviations (std) of these episode durations (**6 features** in total); likelihood of transitioning from one entity to another- *e.g.*, transition

	MPES BOS	OME BOS	AVEID
Unit of Assessment	5-minute observed periods, coded with 0,1 or 2	Identified period of engagement, rated on a 7-point scale	User-specified observed periods of time (flexible granularity)
Attention	Active engagement (Did target activity), Passive engagement (Watched target activity), Other engagement.	Attention intensity (1 denoting *no attention*)	3 raw + 18 derived gaze-based statistics over observed period.
Attitude	Pleasure and anxiety as proportion over observed period.	Attitude valence with (1 denoting *strongly negative*, 3 denoting *neutral*, and 7 denoting *strongly positive* affect).	Proportion of negative and neutral-or-positive affect over observed period.

Table 1. Measuring *attention* and *attitude* via the MPES and OME scales, and the matching measures used with AVEID

from focusing on tablet to focusing elsewhere; this gives rise to **six transition probability features** corresponding to 3 *permute* 2 entity transitions. An additional **six gaze flux features** denoting gaze flux *into* and *out of* the three entities were obtained from marginal likelihoods-- *e.g.*, P(gaze flux *into* tab) = P(fac → tab) + P(others → tab), where → denotes a gaze transition.

To quantify attitude, we computed the proportions of positive (neutral or happy emotion) and negative (angry, sad or disgusted) affect over the observation period from the per-frame facial emotion labels.

In terms of computational hardware, AVEID requires a Graphics Processing Unit (GPU) for fast video processing. The current system is implemented on a Xeon processor with 64GB RAM, and 12 GB NVIDIA GeForce GTX 1080 Ti GPU memory.

BOS FOR VALIDATION
We validated AVEID by comparing the obtained attention and attitude measures against expert annotations acquired for the OME and MPES scales, whose descriptions follow.

The **Observational Measure of Engagement** (OME) [3] is an observational scale to directly assess engagement in PwD. For this BOS, observers are first required to detect *time periods denoting PwD engagement,* and score the attention and attitude levels of the patient within these engagement periods. OME represents a very coarse-grained assessment of PwD engagement, and can best facilitate examination of engagement periods, as no codings are made when the dementia patient is disengaged from the target activity. The **Menorah Park Engagement Scale** (MPES) [2] is a more fine-grained BOS, as PwD engagement is assessed over 5-minute time periods. Three types of engagement, namely, active engagement with the target device/activity, passive engagement and engagement with others, are measured in this BOS (Table 1).

Expert Score acquisition
All annotated videos were as shown in Fig.1, where a PwD engages with an interactive tablet aided by a facilitator [7]. For OME scoring, a dementia care therapist with 10 years experience indicated periods of patient engagement in seven 15-minute video segments according to the following OME definition: "amount of attention the person was visibly paying to the stimulus (tablet) via eye movements;

manipulating/holding and talking about it." [3]. MPES scores were provided by researchers trained to attain 0.8 (Kappa) inter-rater reliability. They scored 5-minute segments from 20 videos (30 minutes each), for active engagement (*Did target activity*), passive engagement (*Watched target activity*) and engagement with other stimuli (*Attention on activity other than target*) on an ordinal scale (Table 1). Table 2 summarizes the annotation statistics.

BOS	No. Videos	No. Video Segments	Annotated by
OME	7	-	Therapist
MPES	20	130	Trained researchers

Table 2. Scores were obtained from experienced therapists or trained researchers.

VALIDATION STUDY RESULTS AND DISCUSSION

Measures of Attention
In the OME BOS, engagement periods are identified and then annotated for attention level and attitude. So, we computed the proportion of patient's gaze focus on the target activity space for a) those segments where the therapist indicated engagement, and b) the remaining video segments where the therapist inferred disengagement. Fig. 3 presents the computed gaze proportions for seven videos. Higher distribution of gaze focus on tablet was clearly noted during engagement periods, as confirmed by a two-sample Kolmogorov-Smirnov test at *p*<0.001. Therefore, gaze focus on the target activity space was sufficient to convey the notion of attention as with the OME.
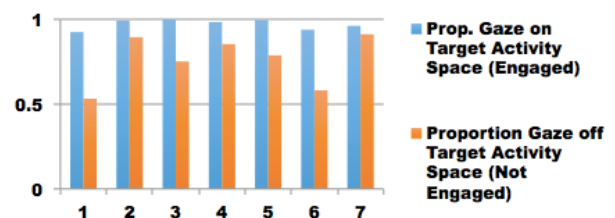

Figure 3 Gaze proportions on tablet (engagement device) during *engaged* and *not-engaged* periods as per OME BOS.

The MPES BOS quantifies attention over 5-minute intervals. Fig. 4 presents Pearson correlations computed between active, passive and other engagement MPES scores, and the 21 AVEID attention features in the form of a 3×21 grayscale image. Negative and positive correlations

are respectively denoted by darker and lighter shades. Red and cyan symbols respectively denote significant (*p*<0.05) and marginally significant (*p*<0.1) correlations.
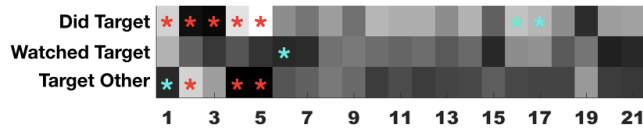

Figure 4. Correlations between gaze-based AVEID features and MPES scores for active (Did Tgt), passive (Watched Tgt) and other engagement (Tgt Other). Red and cyan marks denote correlations significant at (p<0.05) and (p<0.1).

Active engagement is significantly and positively correlated with the extent of gaze focused on tablet, mean duration of tablet gazing episodes and std of these episodic durations, while being negatively correlated with the extent of gaze focus on the caregiver and other areas. Active engagement is also marginally and positively correlated with the gaze flux in and out of the target activity area, suggesting that focus on target activity area as well as periodic gaze switching when communicating with the facilitator are linked with higher active engagement scores as assessed by the expert.

On the other hand, passive engagement marginally and negatively correlates with the gazing durations on the facilitator. This pattern is concurrent with the MPES description of passive activity where the PwD behaves with less enthusiasm and is not having social interactions with the facilitator. The final MPES item, engagement with other, positively correlates with gazing on facilitator, and negatively with the mean and standard deviation of gazing episode durations on the target activity area. This suggests that engagement with other activities, as coded by the experts, is associated with behaviors where the PwD directs attention more toward the facilitator rather than toward the presented activity.

To summarize, gaze on target as a correlate of engagement is validated by both the OME and MPES coding methods. Furthermore, attention measures computed via AVEID are able to capture a number of aspects concerning these BOS.

Measures of Attitude
Inferring facial emotions of PwD is a known challenge due to older adults exhibiting facial emotions in a controlled manner, ageing skin and muscles significantly modulating facial appearance, and indicating a prominently negative affect [6]. Additionally, PwD often display flattened affect [9]. However, greater engagement from the PwD should also elicit a positive reaction from the facilitator whose role is to promote such behavior. As facilitators are neurotypical adults whose facial emotions can be better recognized with available computer vision tools, we therefore examined if higher MPES attitude scores correlate better with the facilitator's facial emotions (Fig. 6). This hypothesis turned out to be true, with pleasure scores correlating significantly and positively with the proportion of positive facial affect

exhibited by the facilitator (*r*=0.24, *p*<0.01). Therefore, examining facilitator behavior could provide crucial cues for measuring engagement in PwD.

Overall, results reveal that AVEID can effectively capture patient (and facilitator) behavior indicative of attention and attitude. Also, since AVEID measures are based on *per-frame* gaze and emotion labels, it is possible to go beyond coarse engagement measures that BOS provide. E.g., even though the patient's verbal behavior was not captured in the videos, the frequency with which the patient directs his/her gaze towards the facilitator may serve as an effective cue to this end. Finally, gazing and attitude estimation can be reliably accomplished for small-space activities (where the patient's face is clearly visible), facilitating evaluation of multiple engagement designs.
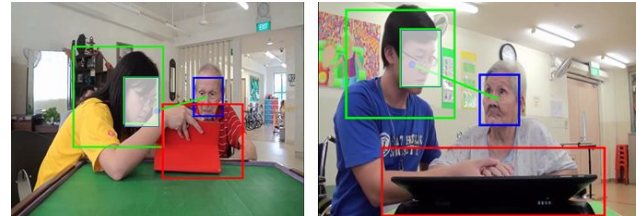

Figure 5 Examples where gaze focus estimation is incorrect (zoom to view).

Figure 6 Exemplar emotion estimation results. Facial emotions of the PwD and facilitator are correctly identified (left). PwD's emotion is incorrectly estimated, but facilitator's emotion is correct (right) (zoom to view).

CHALLENGES AND LIMITATIONS
Figures 5 and 6 illustrate the challenges involved in video-based engagement measurement for PwD, where the observation videos are captured under unconstrained settings. Fig.5 presents two examples of incorrect gaze focus estimation due to the closeness of the facilitator to the tablet, and due to the 2D video information being insufficient to model the 3D world. Likewise, in Fig. 6 (right), the patient's facial appearance is mistaken by the algorithm as an exhibition of sadness. Future work will focus on addressing these challenges.

CONCLUSION
We presented AVEID video-based analytics system that was found to successfully capture various aspects of BOS employed for measuring engagement in people with dementia. Future work will focus on addressing the limitations in our current implementation, and employing additional modalities (such as verbal behavior) for measuring engagement among people with dementia.

REFERENCES

1. 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: 267–275. https://doi.org/10.1016/j.intcom.2010.03.003

2. Cameron J. Camp, Michael J. Skrajner, and Greg J. Gorzelle. Engagement in Dementia. In *Assessment Scales for Advanced Dementia*. HPP, Baltimore, 65–69.

3. 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 : official journal of the American Association for Geriatric Psychiatry* 17, 4: 299–307. https://doi.org/10.1097/JGP.0b013e31818f3a52

4. Natalie C. Ebner, Michaela Riediger, and Ulman Lindenberger. 2010. FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods* 42, 1: 351–362. https://doi.org/10.3758/BRM.42.1.351

5. Stu Favilla and Sonja Pedell. 2013. Touch Screen Ensemble Music: Collaborative Interaction for Older People with Dementia. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration* (OzCHI '13), 481–484. https://doi.org/10.1145/2541016.2541088

6. Mara Fölster, Ursula Hess, and Katja Werheid. 2014. Facial age affects emotional expression decoding. *Emotion Science* 5: 30. https://doi.org/10.3389/fpsyg.2014.00030

7. Pin Sym Foong, Shengdong Zhao, Kelsey Carlson, and Zhe Liu. 2017. VITA: Towards Supporting Volunteer Interactions with Long-Term Care Residents with Dementia. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 6195–6207. https://doi.org/10.1145/3025453.3025776

8. Peiyun Hu and Deva Ramanan. 2016. Finding Tiny Faces. *arXiv:1612.04402 [cs]*. Retrieved October 9, 2017 from http://arxiv.org/abs/1612.04402

9. M. Powell Lawton, K. Van Haitsma, Margaret Perkinson, and Katy Ruckdeschel. Observed Affect and Quality of Life in Dementia: Further Affirmations and Problems. *Journal of Mental Health and Aging* 5, 1: 69–81.

10. Edward L. Mahoney and Diane F. Mahoney. 2010. Acceptance of Wearable Technology by People With Alzheimer's Disease: Issues and Accommodations. *American Journal of Alzheimer's Disease & Other Dementiasr* 25, 6: 527–531. https://doi.org/10.1177/1533317510376944

11. Alex Mihailidis, Scott Blunsden, Jennifer Boger, Brandi Richards, Krists Zutis, Laurel Young, and Jesse Hoey. 2010. Towards the development of a technology for art therapy and dementia: Definition of needs and design constraints. *The Arts in Psychotherapy* 37, 4: 293–300. https://doi.org/10.1016/j.aip.2010.05.004

12. Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning. 443–449. https://doi.org/10.1145/2818346.2830593

13. Viral Parekh, Ramanathan Subramanian, and C. V. Jawahar. 2017. Eye Contact Detection via Deep Neural Networks. In *SpringerLink*, 366–374. https://doi.org/10.1007/978-3-319-58750-9_51

14. Shyamal Patel, Hyung Park, Paolo Bonato, Leighton Chan, and Mary Rodgers. 2012. A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation* 9: 21. https://doi.org/10.1186/1743-0003-9-21

15. Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking? In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett (eds.). Curran Associates, Inc., 199–207. Retrieved October 6, 2017 from http://papers.nips.cc/paper/5848-where-are-they-looking.pdf

16. Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. 2013. Gaze locking: passive eye contact detection for human-object interaction. 271–280. https://doi.org/10.1145/2501988.2501994