The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Predicting Engagement Breakdown in HRI Using Thin-Slices of Facial Expressions

Tianlin Liu
Department of Computer Science
and Electrical Engineering
Jacobs University Bremen
Bremen, Germany
tliu@jacobs-alumni.de

Arvid Kappas
Department of Psychology
and Methods
Jacobs University Bremen
Bremen, Germany
a.kappas@jacobs-university.de

Abstract

In many Human-Robot Interaction (HRI) scenarios, robots are expected to actively engage humans in interaction tasks for an extended period. We consider a successful robot to be alert to Engagement Breakdown (EB), a situation in which humans prematurely end the interaction before the robot had the chance to receive a complete feedback. In this paper, we present a method for early EB prediction using Echo State Networks (ESNs), a variant of Recurrent Neural Networks. The method is based on Action Units (AUs) of human facial expressions. We apply the proposed architecture to a real-world dataset and show that the architecture accurately predicts EB behavior using 30 seconds of facial expression features.

Introduction

For many Human-Robot Interaction (HRI) tasks, we wish to have robots that actively engage humans for an extended period. For example, a competent chatbot should maintain a stimulating conversation for several rounds; a successful educational robot for children should engage children in interaction for a period without distracting them. In general, we consider successful social robots as the ones that not only provide interactions, but actively maintain those interactions.

To maintain an interaction, robots have to be particularly alert to so-called Engagement Breakdown (EB), a situation in which humans prematurely end the interaction, allowing the robot no chance to receive the complete feedback it expects (Ben Youssef et al. 2017). We aim to predict EB at an early stage during an interaction, such that re-engagement strategies can be launched to prolong the interaction time.

In this paper, we present an architecture that predicts EB based on human facial expressions from a narrow window of HRI experience (called thin-slices in Psychology terminology (Ambady and Rosenthal 1992)). Following the Affective Computing approach to affect measurement (D'Mello, Kappas, and Gratch 2017), we build the prediction architecture with two components. In the feature extraction component, we extract the facial muscle movements of humans using Facial Action Coding System (Ekman and Friesen 1978). In the affect estimate component, we use Echo State Networks (Jaeger and Haas 2004), a variant of Recurrent Neural Networks, to model the link between facial expression features and EB behaviors. We evaluate the effectiveness of the proposed architecture on the UE-HRI dataset (Ben Youssef et al. 2017), which is a newly published open-source dataset containing video clips of spontaneous interactions between humans and a humanoid robot. Experimentally, we show that our model accurately predicts EB using only 30 seconds of facial features, up to EB behavior in 10 minutes.

The contributions of this paper are two-fold. First, from the perspective of HRI, we present an EB prediction technique, facilitating the design and implementation of an EB-alert mechanism for robots. Second, from the perspective of affective computing, we empirically validate that there is a computational model that links the affect estimate of EB and machine-readable facial expression features, such that the underlying emotion theory that explains the link could be investigated.

Related Work

Recognized as an essential human response to computer-mediated activities (Laurel 1991), engagement has been widely studied in Human-Computer Interaction and Human-Robot Interaction. Authors of seminal works argue that Engagement has four distinct stages (O'Brien and Toms 2008): Point of engagement, period of sustained engagement, disengagement, and re-engagement. In this paper, we are particularly interested in the second and third stage, i.e., the period of sustained engagement and the disengagement.

There are roughly three perspectives from which to study engagement in HRI (Leite et al. 2015): From the first perspective, researchers examine "which features or social cues robots should be endowed with to increase participant's engagement with the robot" (Sidner et al. 2005; ACM 2014; Vázquez et al. 2014). From the second perspective, researchers investigate how robots can automatically recognize engagement and disengagement (Bohus and Horvitz 2009; Xu, Li, and Wang 2013; Leite et al. 2015; Higashinaka et al. 2016). From the third perspective, researchers aim to "predict engagement or, more importantly, dis-engagement behaviors in real-time, so that the robot or agent can employ repair mechanisms to keep users engaged in the interaction." Our work is in line with the third per-

spective.

Existing automatic engagement and dis-engagement prediction schemes exploit various features, such as gaze (Nakano and Ishii 2010; Bohus and Horvitz 2009), body posture (Leite et al. 2015), voice (Leite et al. 2015), smiles (Xu, Li, and Wang 2013), and gestures (Rich et al. 2010). However, all these features are, arguably, hand-designed and need ad-hoc implementations. In this work, we propose to use Action Units (AUs), a set of fully standardized facial expression features based on anatomic features of human facial muscle. Regarded as a common standard to systematically categorize the physical expression of human emotions (Calvo et al. 2015), AUs could potentially facilitate end-to-end EB prediction with minimal manual intervention.

A very relevant methodology was first proposed in (Jaques et al. 2016) to study bonding, a closely related process to engagement. Similar to our approach, they use thin-slices of facial expressions to train a classifier to predict whether a person will experience bonding up to a delayed time. Different to our approach, however, they train their classifier based on conversations between human participants and their human partners. They then argue that the trained classifier facilitates the design of an intelligent virtual agent (IVA). However, the required training data of inter-human conversation seems to be a restriction to deploying an IVA; when one wants to upgrade the prediction ability of an IVA, one has to recruit human participants to perform new inter-human conversations, analyze the data, and then feed the examined results to the IVA. Our proposed method, on the other hand, does not require training data of inter-human communication. Instead, we train the model directly based on the interaction between humans and robots. This approach gives us the advantage of improving EB prediction ability "on the fly". As the robots collect more and more interaction data during their lifetime, they are able to train themselves to predict EB without the manual loading of new training data.

The primary challenge of our approach lies in the fact that it is generally difficult to infer affective states of humans based on their facial expressions in a reliable way. This challenge has two elements. First, unlike the case of inter-human conversation, when it comes to human-robot interactions, participants are much less expressive in their facial movements, and therefore it is more difficult to extract and analyze their facial expressions. Second, even if it was possible to accurately extract the facial expressions of humans, they are only weakly linked to the affective states of humans. As argued in (Kappas 2003), in many instances, there is no coherence or only limited coherence between facial activation and underlying affective states. To address the first and second challenge, we need an architecture that is robust enough to distill the limited coherence between facial expressions and EB behavior. We show that this can be accomplished by leveraging Recurrent Neural Networks.

Preliminaries

We start by reviewing the main concepts surrounding our task.

Engagement Breakdown Prediction

Engagement Breakdown (EB) is defined as a situation in which humans fail to complete an interaction with a robot and leave the robot before it had the chance to receive full feedback (Ben Youssef et al. 2017). In an EB prediction task, one aims to forecast the likelihood of EB in real time before it happens. Note that EB prediction is far from trivial even for humans; in previous studies, researchers observed that the human prediction of EB in chat dialogues is tantamount to random guessing (Higashinaka et al. 2016).

Affective Computing

Affective computing (AC) (Picard 1997; Calvo et al. 2015) adopts a computational approach to study affect. With the assumption that a link exists between an affect estimate and machine-readable signals, AC makes inferences on the likelihoods of affective states from the signals. The link modeled by AC does not have to be strong; it is enough to assume that a "beyond-chance probabilistic" (Roseman 2011) link exists such that it connects machine-readable signals with affect estimates.

There are two main challenges in AC approaches (D'Mello, Kappas, and Gratch 2017). The first is to obtain affect abstractions (called affect descriptors or features) from raw signals recorded by sensors. The second is to produce affect estimates from the features. In our architecture, we address the challenge of affect abstractions by tracking the Action Units (AU) features of facial muscle movement, and we address the challenge of affect estimation by using Echo State Networks, which will be introduced in a following section.

Thin-slicing

Thin-slicing is a term in psychology describing the ability of humans to predict various objective outcomes in social and clinical psychology based on short observations. Seminal works (Ambady and Rosenthal 1992) have shown that humans are competent in making instant inferences based on limited interaction experiences. For instance, researchers have shown that humans can make consensus judgment on college and high-school teachers based on a brief and silent video in under 30 seconds (Ambady and Rosenthal 1993). Thin-slicing also has implications for the areas of first impressions (Carney, Colvin, and Hall 2007), speed dating (Houser et al. 2007), and deception detection (Albrechtsen, Meissner, and Susa 2009). More recently, (Jaques et al. 2016) has shown that with thin-slices of facial expressions and body language, a classifier can be trained to predict bonding between intelligent virtual agents and their users. Encouraged by previous success, we propose to use thin-slice facial expressions to predict EB in HRI.

Echo State Network

Inspired by the brain's ability to process information, Echo State Networks (ESNs) (Jaeger and Haas 2004) are a variant of Recurrent Neural Networks (RNNs) that process time-dependent information. Previously, ESNs have been successfully used in tasks such as system identification (Jaeger

2003), nonlinear channel equalization (Jaeger and Haas 2004; Boccato et al. 2011), speech recognition (Triefenbach et al. 2010), robot control (Antonelo, Schrauwen, and Stroobandt 2008), and chaotic time-series prediction (Li, Han, and Wang 2012). ESNs enjoy the advantage of being able to process sequential information in a computationally cheap way, giving rise to reliable in-stream EB prediction in less than a minute.

Given an input signal $\mathbf{u}(n) \in \mathbb{R}^{N_\mathbf{u}}$, the update equations for ESN are

$$\begin{aligned} \tilde{\mathbf{x}}(n) &= \tanh\left(\mathbf{W}^{\text{in}}[1; \mathbf{u}(n)] + \mathbf{W}\mathbf{x}(n-1)\right), \\ \mathbf{x}(n) &= (1-\alpha)\mathbf{x}(n-1) + \alpha\tilde{\mathbf{x}}(n), \end{aligned} \tag{1}$$

where $\mathbf{x}(n) \in \mathbb{R}^{N_\mathbf{x}}$ is a vector of reservoir neuron activations and $\tilde{\mathbf{x}}(n) \in \mathbb{R}^{N_\mathbf{x}}$ is its update, all at time step n, $\tanh(\cdot)$ is applied element-wise, $[\cdot;\cdot]$ stands for a vertical vector concatenation, $\mathbf{W}^{\text{in}} \in \mathbb{R}^{N_\mathbf{x} \times (1+N_\mathbf{u})}$ and $\mathbf{W} \in \mathbb{R}^{N_\mathbf{x} \times N_\mathbf{x}}$ are the input and recurrent weight matrices, respectively, and $\alpha \in (0, 1]$ is the leaking rate.

The linear readout layer is defined as

$$\mathbf{y}(n) = \mathbf{W}^{\text{out}}[1; \mathbf{u}(n); \mathbf{x}(n)], \tag{2}$$

where $\mathbf{y}(n) \in \mathbb{R}^{N_\mathbf{y}}$ is network output, and $\mathbf{W}^{\text{out}} \in \mathbb{R}^{N_\mathbf{y} \times (1+N_\mathbf{u}+N_\mathbf{x})}$ is the output weight matrix.

To train an ESN, one follows the general procedure as outlined in (Lukoševičius 2012):

1. Generate a random reservoir $(\mathbf{W}^{\text{in}}, \mathbf{W}, \alpha)$;
2. Use the input signal $\mathbf{u}(n)$ to drive the network according to Equation 1, and then collect the corresponding reservoir activation states $\mathbf{x}(n)$.
3. Compute the linear readout weights $\mathbf{W}^{\text{out}}$ from the reservoir using linear regression, minimizing the Mean Square Error between $\mathbf{y}(n)$ and $\mathbf{y}^{\text{target}(n)}$.
4. Use the trained network on new input data $\mathbf{u}(n)$ computing $\mathbf{y}(n)$ by exploiting the trained output weights $\mathbf{W}^{\text{out}}$.

There are numerous reasonable ways one can use ESNs to carry out classification tasks. We will spell out our designed ESN architecture in more detail in later sections.

Proposed Approach

Problem Setting

We consider human-robot interaction in a open world, where participants are free to interact with the robot and free to leave the robot whenever they want. Given the facial expressions of the human, which are videotaped by the robot's camera, we wish to predict the likelihood of EB based on a short sequence of facial expression features (called thin-slice features). The restriction to a short span is crucial for real applications yet challenging for designing a working architecture. It is crucial in that if the required timespan is too long, say 5 minutes, then it is very likely that the user will leave the robot before it collects enough data to make a prediction, making the early EB prediction effort futile. This is challenging in that our algorithm has to distill informative facial expression features that are coherent to EB/non-EB behavior in an efficient way under severe time pressure. In our problem, we fix the length of thin-slice facial features at 900 video frames, which is 30 seconds in real time.

More formally, we formulate our problem as follows. Suppose we are given episodes of human-robot interaction which are recorded by the sensors of the robot. An episode either ends with an EB (the participant leaves the robot without completing the interaction task) or an NEB (no EB happens, the participant finishes the interaction task). As we are interested in predicting EB/NEB behavior based on thin-slice HRI experience, in the training phase, we subdivide the interaction episode into thin-slices such that each thin-slice contains facial features that last for T frames. Denote one thin-slice as $\mathbf{S} = (\mathbf{I}_1, \cdots, \mathbf{I}_T)$, where $\mathbf{S}$ is the thin-slice, $\mathbf{I}_t$ is a video frame at the time t, and T is the length of the thin-slice. The architecture is supposed to predict the EB or NEB as the aftermath of the interaction episode. Note that an interaction episode might contain many such thin-slices. For each thin-slice in the training sample, we assign a binary classification label (0 for NEB and 1 for EB) to indicate whether EB happens as the interaction ends; this is designed as we aim to predict EB in a delayed time. Our goal is to learn a model $\mathbf{M}$ from the training data, such that in the testing phase, $\mathbf{M}$ takes a new thin-slice of facial feature $\mathbf{S}$ as input and gives out 0 or 1 as output, indicating our hypothesis of NEB or EB up to a delayed time.

We address the problem in three steps. In the first step, we convert the raw videos recorded from the sensor into affective features. In the second step, we train a machine learning model with these affective features. In the third step, with new incoming data, we evaluate the prediction accuracy of the model. A flow chart for these three steps is shown in Figure 1.

From Video Clips to AU Features

When formulating the EB prediction problem in the previous section, we take thin-slices of facial expression features as given. To acquire such features from raw videos, we use the Facial Action Coding System (FACS) (Ekman and Friesen 1978), a fully standardized classification system that codes facial expressions based on anatomic features of human faces. With the FACS, any facial expression can be decomposed as a combination of elementary components called Action Units (AUs). In particular, we use the automated software Emotient FACET, a computer vision program which provides frame-based estimates of the likelihood of 19 AUs. As not all AUs are equally useful for EB prediction tasks, we only maintain N_{AU} number of AUs, $N_{\text{AU}} \leq 19$.

From Affect Descriptors to EB Prediction

In the previous section, we acquired thin-slices of facial expression features, which are multi-channel time series data describing the movements of each AU. With these time series data, we leverage a Recurrent Neural Network architecture to learn the coherence between facial expression and EB behavior. In this work we use Echo State Networks as they

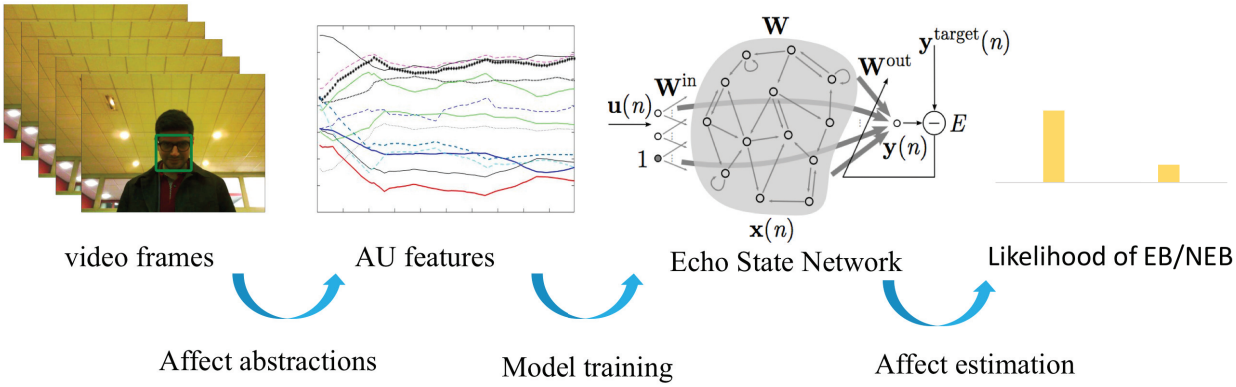

Figure 1: Flow chart of the proposed architecture. We first convert the sequence of video frames into thin-slices of facial expression features. Then we train the ESN with the extracted thin-slices. The trained ESN will make inference on EB/NEB behavior given the new input thin-slices. The ESN illustration is adapted from (Lukoševičius 2012).

can be trained in a computationally cheap way while producing fast outputs such that the EB prediction can be made in-stream.

More specifically, to train the ESN, we first randomly generate the recurrent weight $\mathbf{W}$ and input weight $\mathbf{W}^{\text{in}}$ of Equation 1. We then drive the ESN with input vector $\mathbf{u}^i(t) = \mathbf{S}^i(t,:)$ for $t = 1, \cdots, T$, where each $\mathbf{S}^i(t,:)$ is a vector of AU features at time t of i-th training thin-slice, resulting to the reservoir state vector $\mathbf{x}^i(1), \cdots, \mathbf{x}^i(T)$. We let $\mathbf{q}^i(n) = [\mathbf{x}^i(n); \mathbf{u}^i(n)]$ denote the extended state vector of dimension $N_{\mathbf{x}} + N_{\text{AU}}$, obtained by column-wise concatenate the reservoir state vector with the input vector. We now need to map the information contained in the states $\mathbf{q}^i(1), ..., \mathbf{q}^i(T)$ into EB/NEB hypotheses h_1 and h_2, giving values between 0 and 1 after inputting the i-th thin-slice to the ESN. Let $\mathbf{y}^{\text{target},i}$ be the teacher signal for each training thin-slice i, where $\mathbf{y}^{\text{target},i} = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ if the i-th sample is an EB sample and $\mathbf{y}^{\text{target},i} = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ if it is a NEB sample.

To make our ESN more compact and to boost the training speed even further, we choose a small integer D such that it partitions the T-length time series into $\Gamma = \lfloor T/D \rfloor$ sub-intervals each with length D, where $\lfloor \cdot \rfloor$ is a floor function that rounds down a real number to the greatest integer that is less than or equal to it. For the j-th sub-interval, we only maintain one extended state vector which represents the overall input-reservoir dynamics of the ESN throughout that sub-interval. A natural choice for this extended state vector would be the arithmetic average of all extended states in these sub-intervals (Lukoševičius 2012): Define $\tilde{\mathbf{q}}^i(j) = \sum_{k=(j-1)\cdot D+1}^{j\cdot D} \mathbf{q}^i(k)/D$, $j = 1, \cdots, \Gamma$. We use two output units $(\mathbf{y}_m)_{m=1,2}$, connect each of them to the $\tilde{\mathbf{q}}^i(j)$ by an $1 \times (N_{\mathbf{x}} + d)$ sized output weight vector $\mathbf{w}_m$, and compute $\mathbf{w}_m$ by linear regression of the targets $\mathbf{y}_m^{\text{target},i}$ on all $\tilde{\mathbf{q}}^i(j)$, where $m = 1$ and $2, j = 1, \cdots, \Gamma$. This architecture was first formally introduced in section 4.7 of (Lukoševičius 2012).

	Training (EB + NEB)	Testing (EB + NEB)
Videos	16 (11 +5)	16 (11 + 5)
Thin-slices	96 (38 + 58)	103 (51 + 52)

Table 1: The training and testing data split-up in the UE-HRI dataset. Recordings for training and testing are assigned randomly under the constraint that each training and testing set contains 11 EB and 5 NEB recordings.

Experiments

Dataset Description

To evaluate the proposed EB prediction model, we use the UE-HRI dataset (Ben Youssef et al. 2017), a recently published open-source dataset containing spontaneous interactions between humans and a humanoid robot. The videos of the UE-HRI were collected in a hallway of a university, where participants were free to establish, maintain, and break interactions with the robot. Among the 54 episodes of interactions in the UE-HRI dataset, 32 of them are from mono-user interaction, i.e., the human interacts with the robot on a one-on-one basis throughout the video recording. In this study, we only use the data of mono-user interaction. In these 32 mono-user interaction recordings, 22 users left the interaction before the end (labeled as EB), and 10 users stayed until the end (labeled as NEB). Although for each episode of the interaction the dataset provides a rich stream of hierarchical information, from low-level ones, such as raw video signals, to high-level ones, such as smile-degree, in this study we only use the low-level ones, i.e., the raw videos recorded from the sensors of the humanoid robot.

We use 16 recordings to train our model and the other 16 to test our model. The training and testing set split is performed randomly, with the constraint that each training and testing set contains 11 EB and 5 NEB recordings. After we form the training and testing sets of recordings, we subdivide each recording into thin-slices, each containing 30 seconds of facial expressions. Note that for each interaction record-

Action Unit	Description
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU6	Cheek Raiser
AU10	Upper Lip Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU17	Chin Raiser
AU18	Lip Puckerer
AU20	Lip stretcher
AU24	Lip Pressor
AU25	Lips part
AU43	Eyes Closed

Table 2: The 13 Action Units used in the proposed architecture.

	ESN parameters
No. internal units	1000
No. sub-intervals of extended states	300
Spectral radius of reservoir	0.1
Ridge regression constant	0.1
Leakage	0.2

Table 3: Parameter selection for the ESN model.

ing, the total time of its thin-slices might be much shorter than the length of the recording. This is because not all video frames contain a human face; the recordings start before the interactions and last longer than the interactions. Also, when participants bend over to sign the user agreement (i.e., the agreement to be recorded) and to fill out questionnaires, the video does not contain the face of the participants. We summarize the training and testing data we use in Table 1.

Experiment Setup

Using 30-second facial expression thin-slices as described in the previous section, we are now ready to train the ESN. We first describe the input preprocessing for the facial expression thin-slices. Recall that the Emotient FACET provides frame-based estimates of the likelihood of 19 AUs. After examining the training data, we decide to discard 6 of these AUs because these channels of AUs only sparsely contribute to the facial movement during the interaction. We list the 13 AUs we use for ESN training in Table 2. We smoothen the time series by subtracting its moving local average in the neighborhood of 20 frames. We normalize the data between 0 and 1 as this may help to keep the inputs $\mathbf{u}(n)$ bounded, avoiding outliers (Lukoševičius 2012).

A few hyper-parameters for ESN have to be selected to govern the network dynamics. To select reasonable parameters, we use a stratified 3-fold cross-validation and find the parameters documented in Table 3 perform well in the metric of cross-validation errors. Empirically, we find our ESN is very robust against the jiggling of parameters.

	Proposed method
F1 score	0.76
Accuracy	0.76
Precision	0.74
Recall	0.78
False Negative Rate	0.22
JS Divergence	4.24e-04

Table 4: EB prediction result using thin-slices in the testing set. The numbers are rounded up to 2 fraction digits.

Experiment Results

With the preprocessed input data and selected parameters, we feed all the training thin-slices of facial expression features into the ESN. We train our ESN model using 96 thin-slices from 16 episodes of interaction, and evaluate the prediction result on 103 thin-slices from the remaining 16 episodes . The test results on these 103 thin-slices are reported in Table 4. As suggested by (Higashinaka et al. 2016), we use both classification-related metrics and distribution-related metrics to evaluate our result. We also include the False Negative Rate, a metric emphasized in (Bohus and Horvitz 2009). The metrics we use are listed below.

- Accuracy: The number of correctly classified EB and NEB slices divided by the total number of slices to be classified.
- F1-score, Precision, Recall, False Negative Rate: The F1 score, precision, recall, and False Negative Rate (miss-rate) for the classification of the EBs.
- JS-Divergence: Distance between the predicted distribution of EB/NEB and that of the ground-truth calculated by Jensen-Shannon Divergence.

As our method is a first benchmark result on predicting EB behavior on the UE-HRI dataset, we do not include any other baseline for comparison. However, it is obvious to see that the proposed method significantly outperforms a random decision maker. This result is highlighted as we restrict ourselves to using only 30 seconds of facial features.

We also want to emphasize the training speed of our method. As introduced in the previous section, one advantage of ESN is its computational cheapness in both learning and testing. Our experiment have been carried out using MATLAB (release R2014b for Macintosh) on a 2GHz Intel Core i7 Macintosh notebook computer with 8 GB RAM. The measured CPU-time of learning and testing are 38 and 39 seconds with variance 0.2 seconds, respectively.

Conclusion & Future Work

We considered the task of predicting Engagement Breakdown in Human-Robot Interaction. Inspired by the Affective Computing approach to affect measurement, we designed an architecture that first extracts facial expression features from sensor signals, and then produces affect estimates based on these features using the Echo State Network.

We empirically tested the proposed architecture on the UE-HRI dataset, a newly published open-source dataset containing HRI recordings in an open-world setting. Establishing the first baseline for EB prediction on the UE-HRI dataset, we demonstrated the prediction accuracy and computational efficiency of our architecture.

Future challenges include extending these results to (i) environments with different HRI tasks and possibly with different robots, (ii) environments with multi-party interaction. Another future direction is to verify whether the same features that predict EB in human-robot interaction apply to human-IVA interaction, as is the scenario considered in (Jaques et al. 2016). Furthermore, as (Jaques et al. 2016) has shown that the facial features learned from inter-human interaction provide design implications for an IVA, it is natural to consider if the reverse holds; do the features learned from EB in HRI have any implication on inter-human interaction?

Acknowledgments

We are grateful to Dennis Küster, Arturo Gomez Chavez, and Herbert Jaeger for their helpful input. We appreciate Angelica Lim for alerting us the newly released UE-HRI dataset. We would also like to thank the anonymous reviewers for their comments and recommendations.

References

ACM. 2014. *Social Engagement in Public Places: A Tale of One Robot*.

Albrechtsen, J. S.; Meissner, C. A.; and Susa, K. J. 2009. Can intuition improve deception detection performance? *Journal of Experimental Social Psychology* 45(4):1052–1055.

Ambady, N., and Rosenthal, R. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *psychological bulletin*.

Ambady, N., and Rosenthal, R. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology* 431.

Antonelo, E.; Schrauwen, B.; and Stroobandt, D. 2008. Mobile robot control in the road sign problem using reservoir computing networks. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, 911–916. IEEE.

Ben Youssef, A.; Clavel, C.; Essid, S.; Bilac, M.; Chamoux, M.; and Lim, A. 2017. UE-HRI: A New Dataset for the Study of User Engagement in Spontaneous Human-Robot Interactions. In *ACM International Conference on Multimodal Interaction*.

Boccato, L.; Lopes, A.; Attux, R.; and Von Zuben, F. J. 2011. An echo state network architecture based on volterra filtering and pca with application to the channel equalization problem. In *The 2011 International Joint Conference on Neural Networks (IJCNN)*, 580–587. IEEE.

Bohus, D., and Horvitz, E. 2009. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09.

Calvo, R. A.; D'Mello, S.; Gratch, J.; and Kappas, A. 2015. *The Oxford handbook of affective computing*. Oxford Library of Psychology.

Carney, D. R.; Colvin, C. R.; and Hall, J. A. 2007. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality* 41(5):1054–1072.

D'Mello, S.; Kappas, A.; and Gratch, J. 2017. The affective computing approach to affect measurement. *Emotion Review*.

Ekman, P., and Friesen, W. V. 1978. *Facial Action Coding System: Investigatoris Guide*. Consulting Psychologists Press.

Higashinaka, R.; Funakoshi, K.; Kobayashi, Y.; and Inaba, M. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *LREC*.

Houser, M. L.; Horan, S. M.; Furler, L. A.; et al. 2007. Predicting relational outcomes: An investigation of thin slice judgments in speed dating. *Human Communication* 10(2):69–81.

Jaeger, H., and Haas, H. 2004. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*.

Jaeger, H. 2003. Adaptive nonlinear system identification with echo state networks. In *Advances in neural information processing systems*.

Jaques, N.; McDuff, D.; Kim, Y. L.; and Picard, R. 2016. Understanding and predicting bonding in conversations using thin slices of facial expressions and body language. In *International Conference on Intelligent Virtual Agents*, 64–74. Springer.

Kappas, A. 2003. *What Facial Activity Can and Cannot Tell us About Emotions*. Springer US. 215–234.

Laurel, B. 1991. *Computers As Theatre*. Addison-Wesley Longman Publishing Co., Inc.

Leite, I.; McCoy, M.; Ullman, D.; Salomons, N.; and Scassellati, B. 2015. Comparing models of disengagement in individual and group interactions. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*.

Li, D.; Han, M.; and Wang, J. 2012. Chaotic time series prediction based on a novel robust echo state network. *IEEE Transactions on Neural Networks and Learning Systems*.

Lukoševičius, M. 2012. A practical guide to applying echo state networks. In *Neural networks: tricks of the trade*. Springer. 659–686.

Nakano, Y. I., and Ishii, R. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10. New York, NY, USA: ACM.

O'Brien, H. L., and Toms, E. G. 2008. What is user engagement? a conceptual framework for defining user engagement

with technology. *Journal of the American Society for Information Science and Technology*.

Picard, R. 1997. *Affective computing*, volume 252. MIT press Cambridge.

Rich, C.; Ponsler, B.; Holroyd, A.; and Sidner, C. L. 2010. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Roseman, I. J. 2011. Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emotion Review*.

Sidner, C. L.; Lee, C.; Kidd, C. D.; Lesh, N.; and Rich, C. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*.

Triefenbach, F.; Jalalvand, A.; Schrauwen, B.; and Martens, J.-P. 2010. Phoneme recognition with large hierarchical reservoirs. In *Advances in neural information processing systems*, 2307–2315.

Vázquez, M.; Steinfeld, A.; Hudson, S. E.; and Forlizzi, J. 2014. Spatial and other social engagement cues in a child-robot interaction: Effects of a sidekick. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14.

Xu, Q.; Li, L.; and Wang, G. 2013. Designing engagement-aware agents for multiparty conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13.