In: *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), pages: 1400–1407, Rio de Janeiro, Brazil. IEEE, July, 2018.*

# Learning Empathy-Driven Emotion Expressions using Affective Modulations

Nikhil Churamani, Pablo Barros, Erik Strahl and Stefan Wermter
Knowledge Technology, Department of Informatics
University of Hamburg
Hamburg, Germany
Email: {5churama, barros, strahl, wermter}@informatik.uni-hamburg.de

***Abstract*—Human-Robot Interaction (HRI) studies, particularly the ones designed around social robots, use emotions as important building blocks for interaction design. In order to provide a natural interaction experience, these social robots need to recognise the emotions expressed by the users across various modalities of communication and use them to estimate an internal affective model of the interaction. These internal emotions act as motivation for learning to respond to the user in different situations, using the physical capabilities of the robot. This paper proposes a deep hybrid neural model for multi-modal affect recognition, analysis and behaviour modelling in social robots. The model uses growing self-organising network models to encode intrinsic affective states for the robot. These intrinsic states are used to train a reinforcement learning model to learn facial expression representations on the Neuro-Inspired Companion (NICO) robot, enabling the robot to express empathy towards the users.**

## I. Introduction

Recent years have witnessed a rise in the relevance of social robotics, with researchers working towards building advanced interaction capabilities in robots. As technology becomes more affordable and these robots become more user-friendly and easy to operate, it is expected that people will increasingly use them for assistance in their day-to-day activities. Be it in the form of assistive devices, worktop assistants or even companion robots, the recent boost in computation capabilities has led to these robots becoming increasingly attractive for home domains as well. As conversational partners and companions, these agents are particularly designed to interact with users on a daily basis, not only assisting them in their daily activities but also providing them with a natural and personalised interaction experience. Such robots should be capable of understanding both the user and the environment they operate in, allowing them to become more aware of their surroundings [1] and adapt their behaviour appropriately.

While modelling interactions with humans, emotions and affect analysis play a huge role in interaction design [2]. As emotions are central to any human interaction and are used to communicate intent and motivation [3], social robots should also possess such capabilities [4] in order to understand their users and serve their needs. Over the years, many approaches have been developed for recognising and analysing affect in human conversations. Facial expressions and speech have been examined individually by researchers [5]–[7] to recognise emotions. Furthermore, some studies [8], [9] highlight that emotion recognition benefits from combining multiple modalities. Many recent approaches [10]–[12] thus make use of multi-modal techniques for emotion recognition combining vision, speech and other modalities to improve the overall performance of emotion recognition systems.

In contrast to computational models that take a purely stimulus-driven view on emotions, another approach to affect analysis takes inspiration from how humans view their social partners in order to model improved interaction capabilities in robots. Affective Computing [9] combines computer science research, specifically computational models for signal processing, with psychology and cognitive sciences, where the effect of various psychological concepts [3], [13] and cognitive appraisals [14], [15] is evaluated to equip robots with systems that understand and exhibit emotions. Such an evaluation of affect allows for a holistic view on HRI [2] to be taken into consideration while designing emotionally responsive robots. In this view, emotions are not merely sensory evaluations by the robot but also consider the robot's environment, its subjective experience with the user, individual personality traits and the robot's long-term goals and overall objectives.

To improve their understanding of user behaviour, robots not only need to interpret their spontaneous emotional responses but also the larger context of the interaction. Cooperative robots should model the long-term behaviour [16], [17] of the users, recollecting past interactions with them. This affective model of the robot's interaction with the user or the 'Affective Memory' [18] can be used to empathise with the user, allowing the robot to formulate an appropriate response.

Furthermore, it is understood that humans, while interacting with robots, tend to anthropomorphise them, treating them the same way as they would treat other humans [19]. This allows them to relate better to the robots they are interacting with. As emotions form an important component in human interaction to convey intent and meaning [3], a robot which is able to estimate an emotional state for itself and convey it using expression capabilities [20], [21] will serve as a more natural social companion for humans. Also, grounding long-term and slow-evolving emotional concepts [13] such as *moods* and

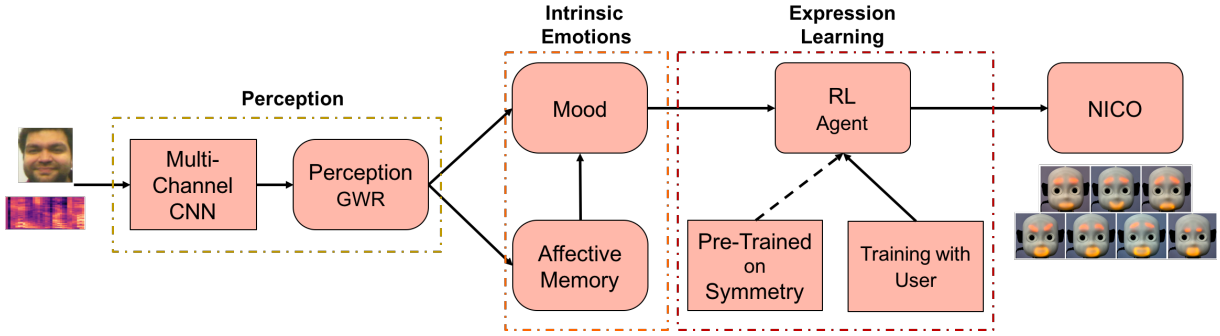



Fig. 1: Perception: Multi-Channel CNN and GWR for feature extraction and representation; Intrinsic Emotions: Mood and Affective Memory GWRs for modelling affect; Expression Learning: RL model for expression generation on the NICO robot.

*attitudes* [2] in robots shall allow them to adapt to the user's behaviour. Compared to spontaneous models of perception which consider only instantaneous stimuli, such long-term emotional models are able to account for temporal changes in stimuli as well as the effect of such changes on the robot's understanding of the environment [21]. These models encode a richer context of the robot's interaction with the user, acting as the intrinsic motivation for robot behaviour. Thus, rather than reacting to spontaneous emotions, the robot takes into account the long-term behaviour of the user to associate a context to the conversation.

Reinforcement Learning (RL) [22] approaches have proven to be successful in learning robot behaviour in different scenarios. In these methods, agents continuously observe the current state of their environment and generate responses. These responses are then evaluated on their aptness and suitability for the current state of the environment, earning the agent a reward. The long-term goal for the robot thus becomes to maximise the reward it attains for its actions, eventually learning the optimum policy for action selection. Interactive reinforcement learning approaches [23] further allow the robots to be trained online while interacting with the user, speeding up the training process. Furthermore, in recent years, actor-critic based approaches [24], [25] have proven successful in learning actions for continuous control, allowing for extremely complex tasks to be solved by robots.

The proposed model (Fig. 1) uses multi-modal emotion perception to develop a growing self-organising model of the affective interaction with the user. It also models an affective memory [21] of the robot's interaction with the user, which is used to modulate the formation and evolution of the intrinsic *mood* of the robot, taking into account the complete interaction with the user. At any given moment, the *mood* models the emotional state of the robot which is used to generate an expression representation, communicating the same. This study explores the facial expression capability of the Neuro-Inspired COmpanion (NICO) robot [26] to express different emotions. It builds on previous works in this direction [27] by not only estimating dynamic models of affect but also learning dynamic and continuous representations of facial expression on the NICO robot, rather than using fixed representations.

## II. Proposed Model

Empathy is considered a key property that a social robot should possess in order to carry out engaging and meaningful interactions with humans [28]. In HRI scenarios, robots that empathise towards the emotional state of the user and adapt their responses accordingly are perceived 'friendlier' by the users [29]. By empathising with the users, the robot can adapt its behaviour best to the emotional state of the user, enriching the user's experience with the robot.

This paper proposes a deep, hybrid neural architecture for learning and adapting robot behaviour based on its affective appraisal of an interaction. The model consists of three modules, each of which focusses on different aspects of affect modelling and behaviour planning using the NICO robot. The Emotion Perception module (Section II-A) proposes a multi-modal approach for emotion perception and uses a growing self-organising model to represent the emotional stimulus. The Intrinsic Emotion module (Section II-B) uses this representation to model long-term and slow-evolving models of affect, providing the emotional appraisal at any given time. The Expression Learning module (Section II-C), finally, learns to express the emotional state of the robot using the facial expression capabilities of NICO. The three modules, together, allow the robot to not only perceive the emotions expressed by the user but also form a long-term model of the user's affective behaviour, estimate an emotional state for itself in the form of its *mood*, and express the mood back to the user, making the robot more socially responsive.

### A. Emotion Perception

Humans, while interacting with others, use various verbal and non-verbal cues to express intent and meaning in a conversation [30]. Apart from the content of the conversation, speech intonation and other non-verbal cues, such as facial expressions and body gestures, add to the meaning of the conversation. Thus, social robots, which are designed to work with humans, should be able to interpret these cues while appraising any interaction with the user.

This paper proposes a Multi-Channel Convolutional Neural Network (MCCNN), based on the earlier works of Barros et. al [11], to examine facial and auditory features for emotion classification (Fig. 2). The network consists of two channels

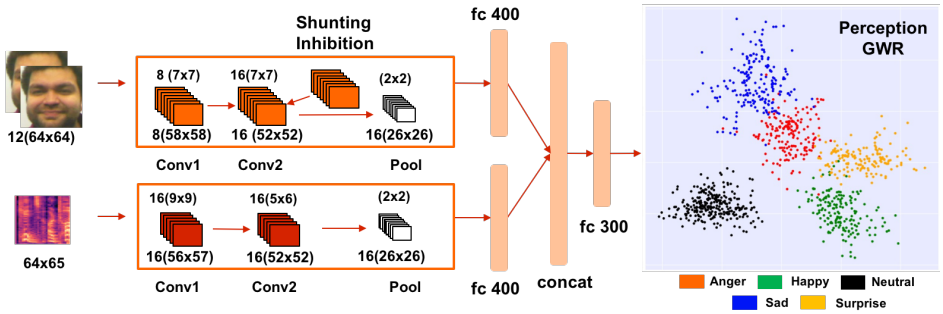



Fig. 2: Multi-Channel Convolutional Neural Network (MCCNN) used to train a Growing-When-Required (GWR) Network (shown here using the two principal components, with the colour representing the emotions) for Feature Representation.

for processing visual and auditory information separately and then combines the learnt features into a single dense representation. For this study, a subset of five emotions, namely *Anger, Happiness, Sadness, Surprise and Neutral*, were chosen as NICO was shown to express these five emotions unambiguously to the users [26], [27].

The vision channel takes as input, the mean of every 12 frames of face images (RGB) of size $(64 \times 64 \times 3)$ from the NICO camera, which operates at 25 frames per second. The faces are extracted using the pre-trained frontal face detector model from the Dlib[1] python library. This input is then passed through a convolutional layer consisting of 8 filters with a filter size of $(7 \times 7)$ convolving over the mean face image. This is passed to another convolutional layer consisting of 16 filters of size $(7 \times 7)$, followed by $(2 \times 2)$ max-pooling filters. The second convolutional layer uses shunting inhibition [31] resulting in filters that are robust to geometric distortions [11]. The convolutional layers are followed by a fully connected layer consisting of 400 neurons which form a dense representation of the facial features needed for emotion classification. The vision channel is pre-trained using the FABO [32] dataset.

The audio channel, on the other hand, uses Mel-spectrograms computed for each half second of audio signal resampled to 16000 Hz and then pre-emphasised. A frequency resolution of 1024 is used, with a Hamming window of 10ms, generating a Mel-spectrogram consisting of 64 bins (Mel-coefficients) with 65 descriptors each. The audio channel consists of two convolution layers with 16 filters each. The first convolution layer uses a filter size of $(9 \times 9)$ while the filter size for the second layer is $(5 \times 6)$ followed by a $(2 \times 2)$ max-pooling layer. The convolutional layers are followed by a fully connected layer with 400 neurons forming a dense representation for the auditory features. The audio channel is pre-trained on the SAVEE [33] dataset.

The dense layers from both the vision and audio channels are concatenated into a dense layer of 800 neurons and fed to a fully connected layer consisting of 300 neurons which encodes a combined dense representation for both visual and auditory features. The MCCNN is then trained by loading the pre-trained vision and auditory channels and training the whole network together using the audio-visual samples from the SAVEE [33] and RAVDESS [34] datasets. The hyper-parameters for the entire network were optimised using the Hyperopt [35] library achieving an accuracy score of 0.74.

Although the classification labels from the MCCNN enable the robot to evaluate the spontaneous emotion expressed by the user, to allow for a developmental emotion perception mechanism [21], a robust approach is needed for emotion representation, accounting for the variance with which different users express the same emotions. This is achieved using a Growing-When-Required (GWR) [36] network which incrementally builds the knowledge representation as it gets different inputs by adding or removing neurons based on the activations of the existing neurons. This allows the model to grow whenever the existing neurons are not sufficient to represent the input, accounting for the variance in the stimuli [18]. This can be seen in Fig. 2 where the Perception GWR is able to represent different emotions as different clusters. The Perception GWR network is trained in an unsupervised manner using the dense layer activation (300-$d$ vectors) from the MCCNN network for the SAVEE [33] and RAVDESS [34] dataset. The GWR is trained for 50 epochs with a maximum age of 50 for each neuron.

### B. Intrinsic Emotions

*1) Affective Memory:* Adopting a long-term view of its interaction capabilities, the architecture needs to form a memory model of affect which grows and adapts based on the user's interactions with the robot. In the proposed model, the Affective Memory (Fig. 3) is modelled as a GWR network that learns to adapt to the user's affective behaviour over the time span of the interaction. As the user interacts with the robot, the emotion expressed by the user is recognised and represented by the MCCNN + Perception GWR network (Section II-A). These spontaneous representations are used to train the robot's affective memory, which represents its recollection of the interaction with a particular user [18], [21]. This memory is used to modulate how the current stimulus is evaluated by the robot. For the study, the affective memory is trained using the two winner neurons i.e. the Best Matching Units (BMUs) from the Perception GWR which fire for the input stimulus. The affective memory is trained, for each user, over 50 epochs, with a maximum age of 30 for each node.

[1]http://dlib.net [Accessed 30.11.2017]

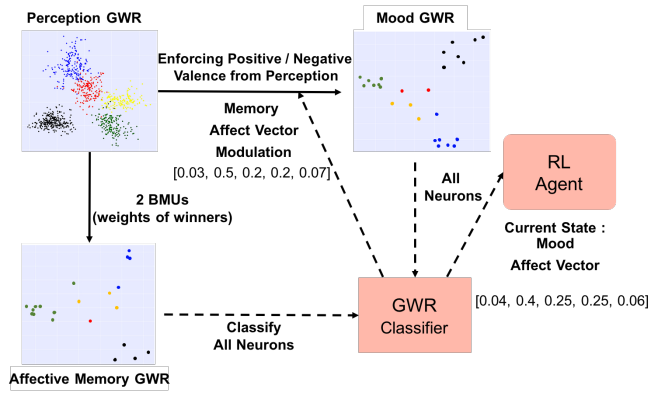



Fig. 3: Perception GWR used to train the Affective Memory which modulates Mood Formation. The current Mood state is used as input for the RL Agent.

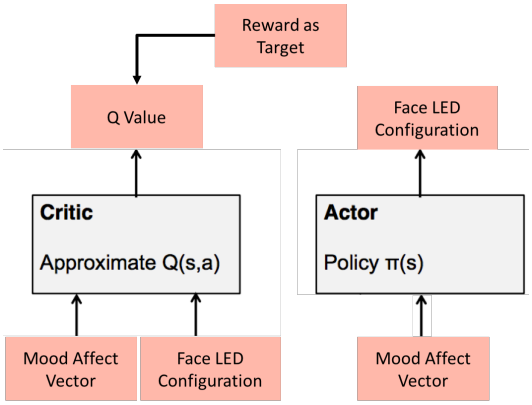



Fig. 4: Actor Critic Learning for emotion expressions using deep deterministic policy gradients.

*2) Mood:* The mood of the robot is estimated as an evaluation of the robot's interaction with the user. It uses the current stimulus, modulated by the affective memory of the user, to estimate an affective appraisal of its environment. Rather than mimicking the user's current expression, it takes into account the behaviour of the user over the entire interaction and uses it to empathise with the user. The mood is also modelled as a GWR network (Fig. 3) making use of its growing and self-organising properties. It is trained using the winner neurons from the perception GWR, positively or negatively enforced using the current state of the affective memory. This modulation (see Section II-B3) allows the robot to use a long-term model of its interaction with the user rather than only spontaneous evaluations. The mood is trained for 50 epochs, with a maximum age of 20 for each node. This allows the mood to slowly evolve during an interaction using the context-encoded information coming from perception.

*3) Affective Modulation:* For the robot to model contextual representations of its interactions with the user, the affective memory encodes the affective behaviour of the user over the entire duration of an interaction. This encoding can be used by the robot as a modulation over the current, spontaneous evaluation of the perception input (Fig. 3). To realise this, the model makes use of an *Affect Vector* notation which represents the state of a GWR at any given instance. To compute this affect vector, all the neurons (input prototypes) from the GWR are classified into one of the five emotions (*Anger, Happiness, Sadness, Surprise or Neutral*). The total fraction of the neurons in the GWR representing each of these emotions is calculated and formulated as a vector notation that represents the current state of the network. For the Affective Memory, this notation (*Memory Affect Vector*) represents the fraction of neurons corresponding to the emotions that the user expressed during the course of the interaction. This vector is used to modulate the current perception input. The winner neurons from the perception GWR are also classified to the encoded emotions and based on whether the *memory affect vector* (representing the Affective Memory) is dominated by positive or negative emotions, the corresponding neurons are enforced from perception by using multiple copies of the prototypes to train the mood. This enforces an input similar to the previous behaviour of the user while lowering the effect of outliers. Similar to the *Memory Affect Vector*, a corresponding *Mood Affect Vector* is also computed by classifying all the neurons from the Mood GWR which represents the current intrinsic affective state of the robot and is used as the motivation for expression generation in the robot.

### C. Expression Learning

For social robots, it is important that they are not only able to perceive and appraise affective interactions but also to express emotions [27]. They need to learn how to express their intrinsic emotional state to the user in order to act as active and natural interaction partners. The facial expressions on NICO [26] are generated using an LED projection system inside the NICO head. This projection system uses two $8 \times 8$ matrix displays for the eyebrows and a $16 \times 8$ matrix for the mouth. To map these matrices to a continuous space, a generator was implemented for facial expressions based on Ricker wavelets represented by a 4-tuple (yStretch, yOffset, xStretch, xOffset). These parameters can be controlled to display different facial patterns on NICO. The robot thus needs to learn the 'correct' combinations of these parameters (each for the two eyebrows and the mouth) in order to generate facial patterns that represent different emotions.

Using the *Mood Affect Vector* (see Section II-B3) as its internal emotional state, the robot needs to learn the correct combination of eyebrow and mouth wavelet parameters to express its mood. One way to do this would be to enumerate all possible combinations of these parameters as potential actions for the robot and use deep reinforcement learning approaches such as Deep Q Networks (DQN) [24] to learn the optimum policy. Since each of the parameters representing a wavelet are real numbers $\epsilon[-1, 1]$, there exists a potentially infinite number of such combinations for each wavelet. Since learning with DQNs in such high-dimensional action-spaces is intractable, the proposed model implements a Deep Deterministic Policy Gradient (DDPG) [25] based actor-critic architecture to learn the optimum policy.

The actor (Fig. 4) receives the current state, in this case, represented by the mood affect vector. It then generates

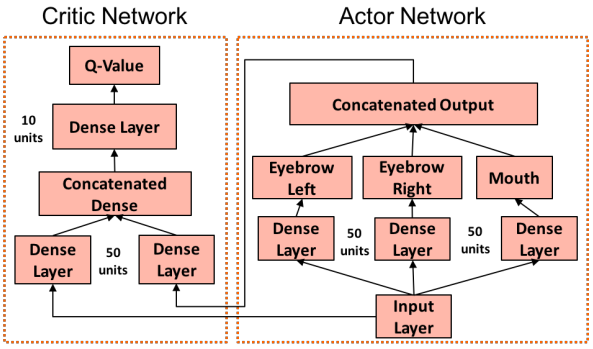
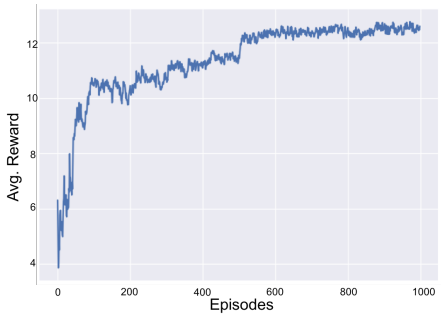


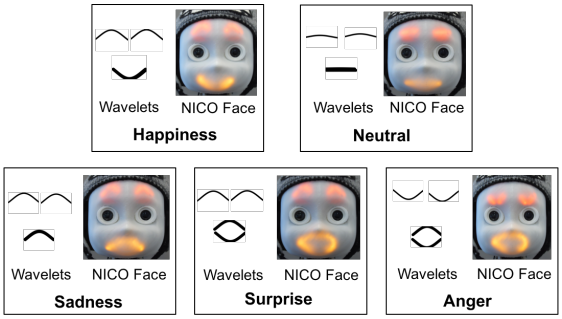

Fig. 5: Architecture for Actor and Critic Networks.

an action corresponding to this state. In this model, the actor generates a 16-tuple representing both the eyebrows (4 parameters each) and the mouth (4 parameters each for upper and lower lip). This action representation, along with the current state is then fed to the critic network, which predicts a Q-value for this state-action pair (Fig. 4). Based on the action generated by the actor, the robot receives a reward from the environment (either using a pre-defined reward function or directly from the user) evaluating the aptness of the action in the given state. This reward is used to estimate the value function for the given state-action pair. The goal of the agent thus becomes to maximise the discounted future reward in order to learn the optimum policy. Since for generating facial expressions, each state is a terminal state, the discount factor ($\gamma$) is set to zero and thus the reward received by the agent at each state acts as the target for the critic network. The reward function is designed to reward symmetry in the wavelets generated by the network. Emotion-specific rules were designed based on the results of previous user studies [26], [27] where users annotated different emotion expressions on NICO. The eyebrow and the mouth wavelets are evaluated individually to assert symmetry, each contributing to the overall reward. The *yOffset* and *xOffset* parameters are reinforced to be close to zero making a wavelet symmetrical along the X and Y axis, while the *yStretch* and *xStretch* parameters are evaluated differently for different emotions based on the facial configurations resulting from previous user studies [27]. The networks are trained off-policy using target networks (slow-tracking copies of the actor and critic networks) to compute the target Q-values at each step. Once the gradients are computed and weights are updated for both the actor and the critic, these target networks are updated using a soft-update rule tracking the learned network. This is done to improve the stability of the learning algorithm. Once the agent learns to produce symmetrical facial representations, it is trained online with users to provide a context and meaning to the learned representations. The users evaluate NICOs expressions in a given situation and reward it depending upon the aptness of the generated expression. This reward replaces the previous reward function and the model is trained by running episodes from a replay buffer storing user interactions.

Both the actor and critic networks are modelled as feed-forward Multilayer Perceptron networks (Fig. 5). The




(a) Average Reward for the actor-critic network over 1000 Episodes.




(b) Network generated wavelets for eyebrows and mouth and corresponding NICO Face LEDs.

Fig. 6: Reinforcement Learning for Facial Expression Generation

actor-network consists of an input layer receiving the 5-tuple *mood affect vector* which is connected to three separate dense layers, each consisting of 50 units, one for each action. Each of these dense layers is then connected to individual output layers, i.e. one for each eyebrow and one for the mouth, which are then concatenated together yielding a single 16-tuple output. The input state (5-tuple) and the output from the actor (16-tuple) are then fed to the critic network. The critic network consists of two parallel dense layers consisting of 50 units each for each of the two inputs which are then concatenated into one dense representation. The concatenated layer is connected to another dense layer of 10 units which is finally connected to a single output unit predicting the Q-value for the given state-action pair.

## III. Experiments and Results

For this study, two different experiments were conducted to train and evaluate the model. The first experiment was conducted offline to pre-train the RL-based expression generator to learn meaningful facial representations using a reward function that rewarded symmetry. The second experiment was conducted with different participants who evaluated the faces generated by the robot depending upon the context of their conversation with NICO. While the offline learning represents social pre-conditioning in the robot as to how different emotions are expressed, the online learning allows the robot to adapt to different users, learning individually tailored responses to their affective behaviour.

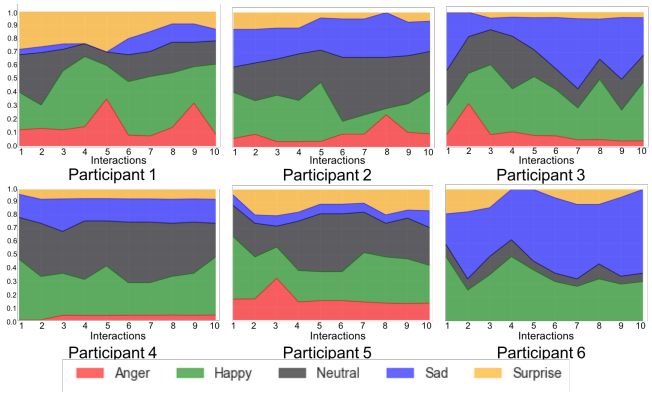



Fig. 7: Robot mood evolving over 10 interactions (each lasting for 5-6 seconds) for each participant. Area under the curves with different colors represent the fraction of neurons in the resultant mood of the robot for each corresponding emotion.

### A. *Offline Training*

For training the model offline, the pre-trained perception module was shown 10-second-long user interaction video clips from the KT emotional dataset [21] with the camera focussing on one participant, who is talking to another. Each clip yielded multiple data samples, each corresponding to half a second of the audio-visual content. The winner neurons from the Perception GWR encoding multi-modal feature representations were used to train the affective memory and the mood GWR. Once the mood was trained, the *mood affect vector* was computed for each video clip, representing the emotional state of the robot as an appraisal of the input stimulus. This process was repeated for all the video clips in the dataset and the corresponding mood affect vectors were saved to train the expression generator model.

A total of 1000 sample mood affect vectors were used to train the reinforcement learning model. Noise was introduced to the data by mixing it with 500 samples generated by randomly drawing values from a normal distribution for each parameter and then passing the resultant vector through a softmax function to return probability values matching the data construct of the mood affect vectors. The model was then trained over 1000 episodes using these 1500 data samples. The average reward received by the model over 1000 episodes can be seen in Fig. 6a.

The offline training resulted in the model learning symmetrical representations for the eyebrows and the mouth for each of the five emotions. The resultant wavelets and the corresponding representations on the NICO face LEDs can be seen in Fig. 6b. As the resolution of the LED matrices on the NICO is not very high (see Section II-C), the learnt representation looked slightly distorted on the NICO robot compared to the simulation wavelets (Fig. 6b). Nonetheless, these were sufficient to express and distinguish between different emotions.

### B. *Online training*

Although the expression representations learnt for the different emotions enable the robot to express the current state of its mood, it also needs to adapt to the user's affective behaviour in any given situation. The robot learns to generate faces which express different emotional states (including mixed representations where two or more emotions dominate robot's mood) but whether these are appropriate in the context of the conversation can only be evaluated by the user. Thus, the second experiment involved the robot interacting with different participants to learn to respond in different social situations. A total of 6 participants were involved in the user study. All the participants were informed about the objectives of the experiment and gave written informed consent for their data to be used for publishing the results of the experiments. After the experiment, the participants were debriefed by the experimenters, who answered any questions arising from their participation in the study.

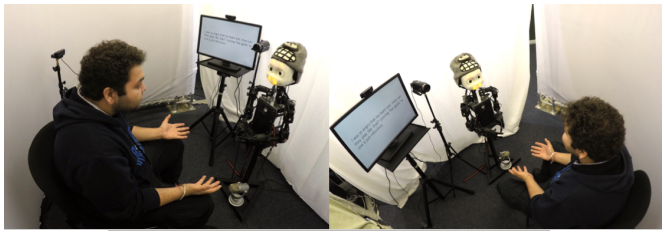


Fig. 8: Experiment Setup: Participant Interacting with NICO.

The participants were requested to sit in front of NICO and interact with it by telling it a story portraying different emotional contexts. This story was split into 21 interaction dialogues. These were presented to the participants using a monitor screen but they were requested to memorise the dialogues and narrate them to NICO while looking at it. The experiment setup can be seen in Fig. 8. To familiarise the participants with the experiment conditions, a trial round was included before the actual experiment. For the further ease of the participants, they were given three practice rounds before each interaction to read and memorise the dialogue and only when they were comfortable with enacting it, their data was recorded. Each dialogue presented 5-6 seconds of audio-visual information to the robot at the end of which the robot generated an expression representation based on the current state of its mood. The participants were asked to reward the robot based on how appropriate they found the generated expression, given the affective context encoded in the dialogue they were enacting. If the robot reacted appropriately, it was given the maximum possible reward, while if it reacted inappropriately no reward was given. Since the faces were generated depending upon the current mood of the robot, which represented the longer context of the interaction, the generated expressions were not limited to the five expressions learnt during the pre-training. For a mixed emotional state (for example, anger mixed with sadness), the model was able to generate novel expression representations representing the mixed state of the mood. This was a significant improvement from the previous studies [27] as facial representations for different emotions were not fixed beforehand.

After training the robot using the story (consisting of 21

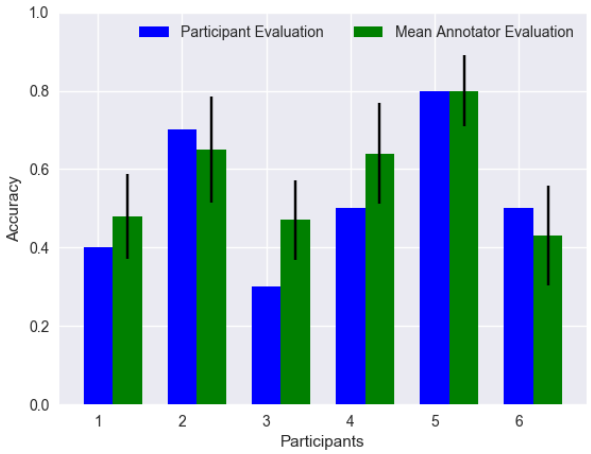



Fig. 9: Participant and Mean Annotator Evaluations ($\kappa = 0.51$) on the appropriateness of the generated expression on NICO.

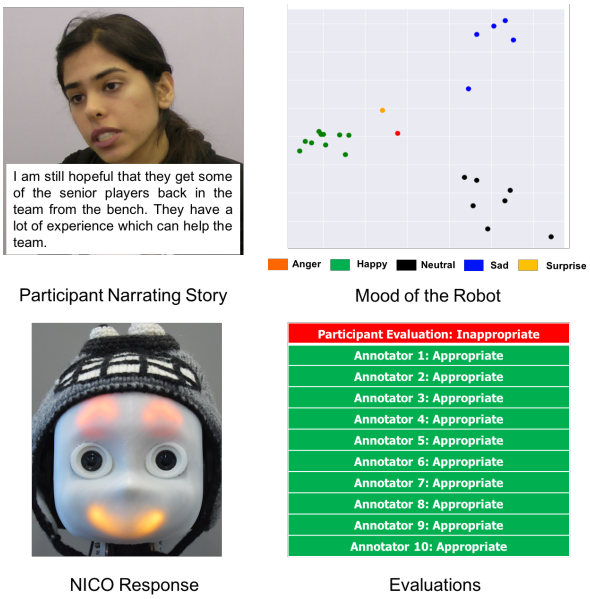



Fig. 10: Top-left: Participant enacting the story to NICO; Top-right: Robot Mood for the current interaction; Bottom-left: NICO responding to the interaction; Bottom-right: Participant and annotator evaluation for the current interaction.

interaction rounds), the participants were asked to evaluate the robot's learning. For this purpose, the participants were asked to enact another story (split into 10 interactions) to the robot. For each interaction, the robot generated a facial expression based on its current mood (see Fig. 7), which the participants annotated to be 'appropriate' or 'inappropriate'. Apart from the participant evaluation, the interactions, consisting of the participant telling the stories and NICO reacting to them, were also shown to 10 independent annotators who were asked to evaluate the robot on the aptness of the generated facial expressions. The results from the participant and annotator evaluations for different participants can be seen in Fig. 9.

## IV. Discussion

This paper presents a deep neural architecture consisting of hybrid models for emotion perception, intrinsic emotions and behaviour generation. The proposed architecture is able to perceive the affective context of a conversation and use it to model long-term emotional concepts such as affective memory and mood. The robot's affective memory of the user acts as a modulation on the perception which influences the robot's mood at any given moment, representing its emotional appraisal of the user's behaviour. The robot uses this intrinsic emotional motivation to interact with its environment by generating a facial expression, thus, responding to the user. The results from the offline learning (Fig. 6a) show that the robot was able to learn to generate facial expression representations based on its current emotional state.

To improve the robot's ability to interact with the user, it needs to adapt to the user's way of expressing emotions. To achieve this, previous works [27] explored interactive learning with users where the robot learns to associate a fixed set of facial expressions to corresponding emotional states. Even though this was shown to be successful in learning emotion expression, fixing expression representations prevented the robot from expressing emotional states which consisted of more than one underlying emotion. This paper addresses this gap by exploring a continuous representation of expression on the NICO robot using the complete face LED matrix to generate expressions. This enables the robot to express more dynamic and mixed emotional states. Although the robot is trained to learn facial expressions, the model could be easily extended to include other modalities such as gesture and speech, to express the emotional state of the robot.

The results from the user study show that the robot was able to associate the learnt expressions with the context of the conversation. The independent annotators rated the system higher in performance as compared to the participants themselves. Such an agreement ($\kappa = 0.51$) attempts to give an objective appraisal [37] of the robot along with the subjective participant ratings. For some participants (for example Participant 2 and 4), both the participant and the annotators agree that the robot responded appropriately to the audio-visual stimuli (see Fig 9). Yet, for some participants, there was a disagreement between the annotators and the participant. One such example can be seen in Fig. 10 where the participant considers the robot's response inappropriate but all the annotators voted otherwise. One explanation for this could be that although the participants intend to express a particular emotion (guided by the story), their expressions convey a different emotion. This is captured by the annotators who use only the emotional expression of the participants to evaluate the robot's responses and not their intentions.

## V. Conclusion and Future Work

Social robots need to adapt their behaviour based on the social context in which they operate as well as the affective behaviour of the user. This paper proposes slow-evolving models of affect to learn expression representations which help the robot to empathise with the user. The experiments offer promising results, with the robot able to generate facial

expressions for complex emotional states, allowing it to participate in the interaction in a more natural way.

Currently, the model takes into account only the resultant affective memory for a particular interaction to modulate the mood of the robot but does not take into consideration instantaneous changes in the affective appraisal to update the mood of the robot. It would be interesting to use multiple influences such as the interaction time and task-specific performance of the robot to estimate the robot's mood. Furthermore, extending the expression synthesis to multiple modalities should allow the robot to interact more naturally and fluidly with users.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge partial support from the German Research Foundation (DFG) under project CML (TRR 169). The authors also thank Matthias Kerzel for the discussions on reinforcement learning in the continuous domain.

## REFERENCES

[1] C. L. Breazeal, "Sociable Machines: Expressive Social Exchange Between Humans and Robots," Dissertation, Massachusetts Institute of Technology, 2000.

[2] R. Kirby, J. Forlizzi, and R. Simmons, "Affective social robots," *Robotics and Autonomous Systems*, vol. 58, no. 3, pp. 322 – 332, 2010, towards Autonomous Robotic Systems 2009: Intelligent, Autonomous Robotics in the UK.

[3] A. Damasio, *Descartes' Error: Emotion, Reason and the Human Brain*. Random House, 2008.

[4] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, Sept 2003.

[5] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, January 2009.

[7] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Face and Gesture 2011*, March 2011, pp. 827–834.

[8] N. Sebe, I. Cohen, and T. S. Huang, "Multimodal emotion recognition," *Handbook of Pattern Recognition and Computer Vision*, vol. 4, pp. 387–419, 2005.

[9] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[10] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April 2011.

[11] P. Barros and S. Wermter, "Developing crossmodal expression recognition based on a deep neural model," *Adaptive Behavior*, vol. 24, no. 5, pp. 373–396, 2016.

[12] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5185–5189.

[13] K. R. Scherer, "Psychological models of emotion," *In Joan C. Borod (Ed.), The Neuropsychology of Emotion*, pp. 137–162, 2000.

[14] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state." *Psychological Review*, vol. 69, no. 5, p. 379, 1962.

[15] R. S. Lazarus, J. R. Averill, and E. M. Opton, "Towards a cognitive theory of emotion," *Feelings and Emotions*, pp. 207–232, 1970.

[16] C. D. Kidd and C. Breazeal, "Robots at home: Understanding long-term human-robot interaction," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2008, pp. 3230–3235.

[17] P. Baxter, T. Belpaeme, L. Canamero, P. Cosi, Y. Demiris, V. Enescu, A. Hiolle, I. Kruijff-Korbayova, R. Looije, M. Nalin *et al.*, "Long-term human-robot interaction with young users," in *IEEE/ACM Human-Robot Interaction 2011 Conference (Robots with Children Workshop)*, 2011.

[18] P. Barros and S. Wermter, "A self-organizing model for affective memory," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 31–38.

[19] B. R. Duffy, "Anthropomorphism and Robotics," in *Symposium of the AISB Convention – Animating Expressive Characters for Social Interactions*. The Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2002, pp. 21–26.

[20] A. Lim and H. G. Okuno, "The mei robot: towards using motherese to develop multimodal emotional intelligence," *Autonomous Mental Development, IEEE Transactions on*, vol. 6, no. 2, pp. 126–138, 2014.

[21] P. V. Alves de Barros, "Modeling affection mechanisms using deep and self-organizing neural networks," Ph.D. dissertation, Universität Hamburg, 2017.

[22] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 1998.

[23] F. Cruz, S. Magg, C. Weber, and S. Wermter, "Training agents with interactive reinforcement learning and contextual affordances," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 4, pp. 271–284, Dec 2016.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013.

[25] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *CoRR*, vol. abs/1509.02971, 2015.

[26] M. Kerzel, E. Strahl, S. Magg, N. Navarro-Guerrero, S. Heinrich, and S. Wermter, "NICO – Neuro-Inspired COmpanion: A Developmental Humanoid Robot Platform for Multimodal Interaction," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 113–120.

[27] N. Churamani, M. Kerzel, E. Strahl, P. Barros, and S. Wermter, "Teaching emotion expressions to a human companion robot using deep neural architectures," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. IEEE, May 2017, pp. 627–634.

[28] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, April 2013.

[29] A. Paiva, I. Leite, and T. Ribeiro, "Emotion Modelling for Social Robots," *The Oxford Handbook of Affective Computing*, pp. 296–308, 2014.

[30] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[31] Y. Frgnac, C. Monier, F. Chavane, P. Baudot, and L. Graham, "Shunting inhibition, a silent step in visual cortical computation," *Journal of Physiology-Paris*, vol. 97, no. 4, pp. 441 – 451, 2003, neuroscience and Computation.

[32] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, 2006, pp. 1148–1153.

[33] S. Haq and P. J. Jackson, "Multimodal emotion recognition," *Machine Audition: Principles, Algorithms and Systems*, pp. 398–423, July 2010.

[34] S. R. Livingstone, K. Peck, and F. A. Russo, "Ravdess: The ryerson audio-visual database of emotional speech and song," in *The 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBCS)*, 2012.

[35] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proceedings of the 12th Python in Science Conference*, 2013, pp. 13–20.

[36] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Networks*, vol. 15, no. 8, pp. 1041 – 1058, 2002.

[37] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Family Medicine*, vol. 37, no. 5, pp. 360–363, May 2005.