Towards individualized affective human-machine interaction

Kazumi Kumagai\textsuperscript{1}, Daiwei Lin\textsuperscript{2}, Lingheng Meng\textsuperscript{2}, Alexandru Blidaru\textsuperscript{2}, Philip Beesley\textsuperscript{2}, Dana Kulic\textsuperscript{2} and Ikuo Mizuuchi\textsuperscript{1}

Abstract—Robots and other autonomous systems interacting with humans should customize their behaviour to their human partner’s preferences. We propose a method for learning and generating robot movement customized to individual preferences. Within a reinforcement learning framework, we generate rewards based on facial expressions observed during the robot’s motion. Robot motions are parametrized; the rewards are used to modify these motion parameters using Q learning. The proposed approach is evaluated in a user study, using an interactive kinetic sculpture. The system interacts with participants and evolves its motion based on the rewards estimated from the participants’ facial expressions. Our results show that, for a subset of participants, the system was able to successfully generate actions that resulted in higher than random rewards. The ability to successfully generate high-reward actions depends on: being able to recognize positive affect from the face, being able to generate actions that are pleasing to the participant, and being able to learn the mapping from rewards to actions.

I. INTRODUCTION

As robots and other autonomous systems increasingly operate in human environments, they will need to adapt to, and learn to accommodate individual preferences of their human partners. The long term objective of our research is to construct systems that can refine and adapt their behaviour based on user preferences that can be observed through natural interaction. In this paper, we propose a method of generating robot behavior that is adapted to each user based on their observed facial expression during human-robot interaction.

Understanding and conveying affect has been found to help robots to behave socially \cite{1,2}. Human facial expressions are one of the keys to estimating emotion \cite{3}. For example, if the facial expression of a human can be observed and used to estimate the affective state of the human, the estimated affective state can help a robot or agent to generate appropriate behavior (e.g. comforting when a human is crying) before a robot starts its behavior. On the other hand, emotions that are estimated after/during robot behavior could be used to estimate the human’s assessment of the robot’s performance and their personal preferences (what kind of behavior is good or not). If a robot could automatically identify each individual’s assessment and personal preferences for its behaviors, and use these assessments to adapt its behaviours to the user’s preferences, the robot designer would not have to prescribe various behaviors to cover a large variety of possible user preferences. Furthermore, the ability to gradually individualize responses to each user could increase the user’s attachment to the robot and foster its long-term use \cite{4}. For example, by learning and customizing the robots behavior based on the users affect, tutor robots\cite{5} or robot mediators \cite{6} for autistic children can learn to provide guidance appropriate for each child. Also, adapting the robots behavior based on the users affect can help therapy robots (like paro\cite{7}) learn appropriate motions that the user prefers.

In this paper, we propose an approach for learning user preferences during on-line human-robot interaction. The robot’s actions are parameterized, and the robot observes the human partner’s facial expression during interaction. A reward function is formulated to reward the robot when facial expressions with positive affect are observed. During the interaction, we adapt the action parameters to maximize reward using reinforcement learning. We also develop an evaluation framework that allows us to test the quality of the facial expression and reward estimate.

This paper is organized as follows: In Section II, we introduce related work on robot action adaptation and individualization. In Section III, we first describe the interactive system we use in our experiments and the system’s perceptual and action capabilities. Next, in Section IV, the proposed method is described. In Section V, we describe the user study and experimental protocol used to evaluate the proposed method. The experimental results are presented in Section VI and discussed in Section VII. Section VIII concludes the paper.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Proposed system for learning and generating appropriate motions based on the observed facial expression $o(\tau)$. $Q(s,a)$ is the action-value function representing the relationship between the state $s$, motion parameters $a$ and the discounted total reward. The function $f_r(\cdot)$ generates the reward $r$ based on the observed facial expression $o(\tau)$.}
\end{figure}

\*This work was supported by Building of Consortia for the Development of Human Resources in Science and Technology and the Living Architecture Systems Group, funded by the Social Sciences and Humanities Research Council of Canada.

\textsuperscript{1} K. Kumagai and I. Mizuuchi are with Department of Mechanical Systems Engineering, Tokyo University of Agriculture and Technology, Koganei, Tokyo Japan \{ kumagai, ikuo \}@mizuuchi.lab.tuat.ac.jp

\textsuperscript{2} D. Lin, L. Meng, A. Blidaru, P. Beesley and D. Kulic are with the University of Waterloo, Waterloo, ON, Canada dana.kulic@uwaterloo.ca
Personalized or individualized robots are those that adapt their behaviors based on preferences of individuals [8]. Such capabilities are expected to be important for partner robots, robot tutors, therapy robots and many others. Personalized or individualized robotics can be divided into two categories: 1) Studies that analyze the effects of robot personalized behaviors on their users and 2) Studies that develop automatic systems for personalizing or individualizing behaviors of a robot.

Regarding the first type of studies, a number of previous works [9][10][11] investigate the effects of personalized/individualized behaviors on human participants in various tasks, using remote control of the robot in Wizard of Oz (WoZ) paradigms. For example, Nikhil et al. [12] survey personalized dialog generation based on the user’s name or preferences, and show that users find personalized robots more familiar or intelligent. In the studies of Leyzberg et al.[9], tutorials given by a robot whose behavior is personalized to individual students’ strengths or weaknesses encourage students to improve their performance. Abdollahi et al.[10] survey the relationship between the elderly and robots that understand user emotions and personalize their interaction during dialog or games. Lee et al.[11] show that the relationship between users and robots becomes stronger when robots generate personalized behavior based on memories.

Considering studies of the second type, Kumagai et al. proposed a method of selecting an appropriate behavior based on facial expressions recognized through an interaction with a person [13]. The method uses the action selection method proposed by Baek et al. [14], which leans the relationships between situations, actions and the evaluations given by an individual human. The method allows the robot to select an appropriate behavior according to individual preferences and context, given a discrete set of behaviors. Gordon et al. proposed a method of selecting appropriate behaviors with a policy tuned according to an individual [5]. Their method was implemented on a robot which encourages a child who plays a game. In their study, a set of robot comments are prepared in advance to encourage a child. The comments are selected using reinforcement learning based on the observed facial expression as a reward. Similar to [13], in this work the set of robot actions is discrete and pre-designed and new actions cannot be generated by the system autonomously.

Key challenges in constructing methods of personalizing/individualizing behaviors of a robot include how to estimate the user’s preferences and how to modify behaviour based on the user’s feedback. Studies on estimating human emotion [15] could be used as methodologies for evaluating the robot’s behavior. However, most current studies on estimating human emotion rely on actor generated emotional displays [16], which may be exaggerated and not observed during human–robot interaction. Furthermore, estimating the user’s emotion does not provide information on how behaviors should be modified.

By representing the robot’s behavior as combinations of continuous parameters of motion, the robot’s behavior could be modified by changing the action parameters so as to increase the estimated emotion. Using reinforcement learning, our system learns how to modify its behavior using the affective evaluation of the human estimated from their facial expression. Unlike previous work, our action space is continuous, so that new actions, not pre-implemented by the designer, can also be generated. Furthermore, as the success of the learning system depends critically on the ability to correctly generate rewards from facial expressions, we introduce a new experimental procedure to validate the quality of the facial expression estimation and subsequent reward.

III. INTERACTIVE SYSTEM

To illustrate the proposed interactive learning system, we use the “Living Architecture System” [17]. A working physical model of the proposed interactive learning system was created by adapting components from an integrated test-bed system currently under development by the Living Architecture Systems (LAS) group [17]. LAS testbeds are interactive environments that aim to engage human observers by combinations of motion, light and sound. The components used for this study focused on motion. Previous implementations of Living Architecture Systems used a method of generating movement and interacting with human visitors based on its internal motivation formulated as a curiosity drive [18][19]. In this research instead we aim to learn the best motion strategy for the interactive system from the user’s affective state during interaction.

The system used for interaction is shown in Fig. 2. The system consists of vše microprocessor-controlled kinetic devices that each contain a flexible armature that contracts and expands. A serrated, feather-like mylar sheet is fitted to each armature. These devices can be individually actuated to generate waving movements. Each kinetic device is fitted with a proximity sensor oriented to receive stimulus from users. The sensor can be used to detect the presence of the user or the users hand nearby. Actions of the kinetic devices are controlled by these sensors and are coordinated to generate a variety of synchronized motions. The device that detects the hand of a participant reaching towards it initiates its motion, followed by motion of its neighboring elements. Our parametrization of the system’s action space is described in Section IV-A.

To interact with the system, the user sits in front of the system shown in Fig. 2, and their facial expression is observed by a USB camera (UCAM-C0220FBNBK) that faces the participant. Using the captured webcam pictures, we use OKAO Vision [20] to recognize facial expressions. The OKAO vision system outputs continuous values of estimated facial expression labels (neutral, happy, surprised, angry and sad face) for each captured picture. The magnitude
of each element is from 0 to 100, and the sum of these five values must be 100.

IV. PROPOSED APPROACH

Fig. 1 illustrates the proposed system for learning appropriate motions based on the observed facial expression. To generate motions which are appropriate for an individual, our method allows robots/machines to modify their motion based on facial expressions observed during human-machine interaction. We first parameterize the motion space using a set of action parameters \( a \). To modify the motions, we use reinforcement learning, with the rewards generated based on facial expressions.

The proposed method follows Algorithm 1. In this Algorithm 1, \( \alpha \) is learning rate and \( \gamma \) is discount factor to update \( Q(s^t, a^t) \).

Algorithm 1 Generate individualized motion

1. **Initialize** \( Q(s, a) \)
2. **for** each episode \( t \): **do**
   1. **state** \( s^t \)
   2. **Select** motion parameter \( a^t = \arg \max Q(s^t, a^t) \)
   3. **Take** motion \( f_a(a^t) \), observe facial expressions \( o^t(\tau) \)
   4. **Calculate** \( r^t = f_r(o^t(\tau)) \)
   5. **Update** \( Q(s^t, a^t) \leftarrow (1-\alpha)Q(s^t, a^t)+\alpha(r^t+\gamma \max Q(s^t, a^t)) \)
3. **end for**

A. State and action spaces

The state vector \( s^t \) contains the perceptions of the system. These could include direct sensor readings or features estimated from the sensor readings. For example, affective states of participants could be estimated from a camera sensor and included in the state vector.

For the system described in Section III, the state vector at each time step \( t \) contains six elements \( s^t = (s_0, \cdots, s_5) \). There are five elements \( (s_0, \cdots, s_4 \in s) \) corresponding to the average value of the observed facial expression, averaged over the window starting when the system first detects a hand in front of any of the proximity sensors, and ending when the system ends its motion. The sixth parameter \( s_5 \in s \) is the index of the proximity sensor which detected the user’s hand.

The action space of the system is represented by combinations of motion parameters which are continuous values represented by the action vector \( f_a(a^t) \). An element of \( a^t \) does not represent one motion, but parameters such as, for example, the velocity of motion, the rotation angle, or the amplitude of movement. This parametrization allows the system to generate a wide variety of motions that may be difficult to generate manually by a fixed library of motions.

Three action parameters were formulated: the direction of motion propagation, the timing of the propagation and the amplitude of motion. The direction of motion propagation was decomposed into three patterns: to the right, to the left and in both directions. The timing of the motion propagation was formulated as a time delay parameter, which controlled how soon after its neighbor each device would initiate its own movement. The amplitude of motion was parametrized by the PWM duty rate of each device’s actuator. In this experiment, the delay time and PWM duty rate were common for all five devices.

The motion parameter vector \( a^t \) at each time step \( t \) contains three parameters to represent the motion of the device. The motion parameter vector elements \( a_i(i = 0, 1, 2) \) represent the value of the PWM duty rate which was input to actuator of the device, the delay time \( \Delta t \) to start its motion and the direction of its motion sequence. The state and action spaces are summarized in Table I.

B. Reward Computation based on Facial Expressions

The observed facial expression is used to allow the system to learn what motion is preferred by each individual. Here, we use the observed facial expression \( o^t(\tau) \) to calculate rewards \( r^t = f_r(o^t(\tau)) \). We investigated two approaches to calculate rewards: one that focused on the average of facial expression (2) and one that focused on the change value of the observed facial expression (2).

\[
  r_{ave}^t = \sum_{i=0}^{T} w_i o_{ave} a_i \quad (1)
\]
\[ r_{del}^t = \sum_i w_{del}^{T(i)} \frac{do_i}{dt} \]  

(2)

Here, \( T \) represents the time window containing the robot’s motion, \( o_i \) represents the type of facial expression (0: neutral, 1: happy, 2: surprised, 3: angry and 4: sad), and \( w_i \) represents the coefficient multiplying each type of facial expression \( o_i \). The values of \( w_i \) were defined so as to be positive/negative if \( o_i \) is positive/negative respectively.

C. Q learning to learn preferences of an individual

Using Algorithm 1, the system learns what motions elicit positive facial expressions from a user. In other words, the system learns a model which represents the relationship between its state and motion parameters and positive facial expressions, which is captured by the action-value function \( Q(s,a) \). Since such a model differs according to each individual, it is difficult to get such a model in advance. Therefore, we use the framework of machine learning to learn the relationship between the state, an appropriate motion and the user’s facial expressions. \( Q(s,a) \) is implemented by a neural network, modeling a continuous function over state and action values. The neural network has three layers (input, output and a hidden layer) with the number of units in the hidden layer equal to half of the sum of the input and output units. During the interaction, the NN is trained by online learning with a set of the sensed state \( s^t \), conducted motion \( a^t \) and updated value \( Q(s^t, a^t) \). This allows the system to generate interpolated estimates for the state-action value function for combinations of \( s \) and \( a \) which have never been experienced before.

The motion parameters \( a \) which maximize the \( Q \) value (policy \( \pi \)) are selected to generate an appropriate motion.

V. EXPERIMENT

To evaluate the proposed method, we conducted a user study where participants interacted with the system described in Section III, using actions generated by the algorithm described in Section IV.

A. Experimental procedure

The study was approved by the Office of Research Ethics at the University of Waterloo. Each study session took approximately 30 minutes. After being informed about the study and providing consent, participants were provided with a brief introduction and an overview of the study. The participant is then seated on a chair in front of a desk, where the living architecture system is placed in front of the participant as shown in Fig. 2. A camera faces the participant and records video of their face and facial expressions throughout the experiment. A total of 17 participants (20-30 years old, 11 males and 6 females) were recruited to interact with the system. Only one participant was interacting with the sculpture at a time.

The participant interacts with the system for 10-15 minutes. During the interaction, the participant can move their hands and arms to interact with the system (Fig. 3). The experiment followed the procedure shown below for each participant.

- The participant answered a questionnaire assessing their affective state before the experiment.
- The participant interacted with the Living Architecture System device in the first condition.
- The participant answered a questionnaire assessing their impression of the shown motion of the device.
- The participant interacted with the Living Architecture System device in the second condition.
- The participant answered a questionnaire assessing their impression of the shown motion of the device.
- The participant answered a questionnaire evaluating the expression on their face during the interaction with pictures which were captured during the experiment.

The two conditions correspond to the two ways of computing the rewards of the system, as detailed in Section IV-B:

- Based on the average of observed facial expression ((1))
- Based on the change of observed facial expression ((2)).

The experiment included two interactions for each condition (Average condition and Delta condition) to enable evaluation of each approach for calculating rewards \( r^t \).

B. Facial Expression Self-Assessment

To check whether the facial expressions were correctly measured and whether reward functions correctly estimated the participants’ positive emotion, we stored sample images from the camera during the interaction. Following the interaction, the participant was shown the collected images and asked to rate their own facial expressions, recorded during the experiment, using the PAD (Pleasure, Arousal, and Dominance) emotional model measurement sheet [21]. Fig. 4 shows the user interface for the facial expression self-assessment questionnaire. Four pictures were shown for each condition. Two were captured when the calculated reward was highest and two were captured when the calculated reward was lowest.
VI. RESULTS

To verify the proposed method, we compared the rewards given when motions were generated in a random manner (random rewards $r^{ran}$) and when using the learned policy (policy rewards $r^\pi$). If the proposed method is successfully generating preferred motions, the policy rewards $r^\pi$ are expected to be higher than rewards following randomly generated motions $r^{ran}$.

Let $r^t \in r(\pi = r^t : t = t_0 \cdots t_{\text{end}})$ be the reward given at time $t$, $R^t$ is calculated by the following formula:

$$R^t = \frac{r^t - \min r}{\max r - \min r}$$

(3)

$\bar{R}^\pi$ and $\bar{R}^{ran}$ are the average values of $R^\pi$ and $R^{ran}$.

A. Recognizing positive affect from the face

In the proposed method, accurate computation of the reward function is based on the estimated facial expression, and therefore the success of learning will depend on the quality of the facial expression recognition.

Fig. 5 shows the correlation between the happy component of the estimated facial expression obtained from the camera and the self-rated facial expression (pleasantness and arousal) for each participant. As the happy emotion is higher in both pleasantness and arousal than neutral, if facial expressions are recognized accurately, we expect to see a positive correlation between the happy component and both self-rated valence and arousal. However, positive correlation is only observed for a subset of participants.

To verify how facial expression recognition affects the generation of rewards, we next calculated $\rho_{e_i,r}$, the correlation between the level of self-rated emotion $e_i$ (pleasant and aroused) and the calculated rewards $r$.

A high correlation indicates that the calculated rewards captured a positive emotion. Fig. 6 and Fig. 7 show the correlation $\rho_{e_i,r}$ for each type of positive emotion. The correlations between (A) $\rho_{e_i,r}$ for each type of emotion (pleasant and aroused) and (B) the difference of rewards $\bar{R}^\pi - \bar{R}^{ran}$ were −0.12 and 0.42 in the Average condition and 0.20 and 0.56 in the Delta condition. Participants whose difference of rewards was positive tend to have higher correlations $\rho_{e_i,r}$.

B. Learning User Preferences during Interaction

Fig. 8 shows the difference between $\bar{R}^\pi$ and $\bar{R}^{ran}$ for each participant.

There were five participants whose policy rewards $\bar{R}^\pi$ were higher than random rewards $\bar{R}^{ran}$ in the Average condition and four participants in the Delta condition. Thus, the system is able to generate improved behaviours for only a subset of users.

C. Example Behaviour Evolution

To illustrate the behaviour of the system when the rewards are higher than random, Fig. 9 shows selected motion parameters during the interaction with P5 in the Average condition. During the first half of the interaction, small values of both motion parameters tend to elicit higher rewards. Then, for
participant P5, the system selected motion parameters that resulted in higher rewards in the latter half of the interaction.

Given that our system only interacted with participants for a short amount of time, the amount of training data is low. Since $Q(s, a)$ is updated based on the observed facial expressions, sensor readings and conducted motion, if a state $s^t$ is one the system has never experienced, the generated motions might not be appropriate, as the learning system may not be able to generalize outside of the training region, especially given the low number of training episodes in our experiment.

Therefore, generated motions are expected to be more appropriate when a state $s^t$ is similar to the states $S^{t-1} = \{s^t|t = t_0, \cdots , t - 1\}$ the system has experienced before.

To estimate how similar a state $s^t$ is to the past states, we calculated the average distances between $s^t$ and subsets of states $S^{t-1}$. Let $s^t$ be a state vector at time $t$, we denote by $d^t$ the average distance between $s^t$ and subsets of states $S^{t-1}$, using the following formula:

$$d^t = \frac{\sum_{i=1}^{t-1} ||s^t - s^i||}{t - 1}. \quad (4)$$

Fig. 10 shows the time transition of rewards $r^t$ and distance $d^t$ for participant P5 in the Average condition. After the first two episodes, the system states are similar to the past states during the interaction.

If learning is successful, we expect the action value function $Q(s, a)$ to increase over time, as the system generates motions which are preferred by an individual and therefore rewarded. To check whether $Q(s, a)$ increases as expected, we calculate the difference of the action-value function from the initial state to each timestep over combinations of all of the states $s$ and motions $a$ by the following formula:

$$Q^t = \sum_{i,j} Q(s^t_i, a^t_j) - \sum_{i,j} Q(s^0_i, a^0_j) \quad (5)$$

Fig. 11 shows the time transition of rewards $r^t$ and distances $d^t$ which is defined the formula (4). When we focus on $t=8$, the reward $r^8$ was low even though the distance $d^8$ was not so large. It means that the learned $Q(s, a)$ was not appropriate in the state $s^8$. Fig. 11 shows that the sum of $Q(s, a)$ changed a lot, which means the mapping of $Q(s, a)$ was modified a lot (Fig. 12) after $r^8$ was given. Then the system was given higher reward $r^9$. At that time, the distance $d^9$ was the smallest ever experienced, and the motion parameters that were not frequently selected before were selected in $t=9$. The system updated $Q(s, a)$ so that the past experience was reflected, showing that the system could update $Q(s, a)$ to favor preferred action parameter settings.

**D. Subjective Evaluations**

After each participant completed the experiment, they were interviewed to obtain their subjective impressions of the interaction. Participants did not perceive a significant difference between the two conditions or that the system gradually modified its motion. Some participants thought
the system was interesting, while one participant thought that some of the motions looked the same and felt it was a bit boring. One participant reported that it was difficult to understand initially what was the motion of the system.

VII. Discussion

In this section, we discuss the factors that influence system performance.

As we can see from the results in Fig. 8, Fig. 6 and Fig. 7, as expected, the performance of the system is highly dependent on the ability to accurately estimate affect from the face. For those participants where affect is more accurately estimated, there is a correlation between either the valence or arousal dimensions and the reward gain. For example, participants P1, P3, P9 and P13 all have positive correlations between arousal and reward gain, and these are the participants for which rewards attained by the system were higher than random.

For a subset of the participants, the motions which were generated based on the proposed method resulted in higher reward than randomly generated motion. The generated motions were novel motions generated by adapting the motion parameters, and differed according to the participants’ preferences. The proposed approach allows the system to explore and refine novel movements that were not pre-specified by a human designer.

Our study has a number of limitations. As discussed above, the quality of the proposed method relies on the quality of facial expression recognition. If the user usually expresses their emotion on their face, the system is expected to gradually learn what kind of motion is preferred by the user and generate an appropriate motion. When the system learns based on the facial detection which includes error, the system will generate inappropriate motions. In that case, if the user demonstrates negative facial expressions, the system may revise the learned mapping based on the negative facial expressions. However, if a user does not show their emotion on their face, this method would require another type of emotion sensor to recognize emotion accurately.

Furthermore, since the duration of the experiment was short, it is not clear how the participants would react during longer interactions. On the one hand, a longer interaction would provide more training data to the system, allowing it to explore a wider range of the action space and better individualize its behaviour to each participant. On the other hand, the participant may become bored with the interaction, or their preferences may change over time. Finally, it is not clear if the system can work if no action is interesting to the user. If there is no motion which is preferred by the user in the possible motion space, another strategy (for example, re-parametrizing the action space to expand the variation of motion) may be required. The representation of the motion space needs to be carefully designed.

The proposed approach may be generalizable to other action spaces in addition to motion, such as, for example, speech. When applying our method to generating robot speech, the sentences or words the robot utters need to be converted to continuous vectors by using word2vec, sentence2vec and so on.

VIII. Conclusions and Future Work

We introduced a method for generating interactive system movement based on on-line facial expression recognition. In the proposed method, the system uses facial expressions as rewards within a reinforcement learning framework to learn motion parameters that generate positive facial expressions in the user. We conducted a user experiment where an interactive sculptural system generated motions based on the observed facial expressions. When we compared motions which were generated in a random manner to those generated by the learning system, for some participants, higher reward motions were shown to be generated through the proposed learning interaction. For those participants whose motions resulted in higher rewards, the calculated rewards corresponded to his/her positive emotions based on his/her facial expression. Since the system modifies its motion through an interaction with an individual based on the observed facial expression, the proposed method is expected to allow the system to generate motion which are preferred by an individual.

In future work, to design a reward function that can help the system to generate good motions, we will study how to estimate human emotion based on facial expression. Facial expression changes as a reaction to the motion may differ.
according to the feature of a motion and action generation function. For example, the facial expression may change slowly when the shown motion is slow and the motion is favored by the user. Such temporal characteristics of facial expression changes should be considered when designing reward functions. We will also aim to implement our method to generate longer and more complex interactions. Finally, other modalities to estimate human response could be used to calculate rewards.

REFERENCES