Impression Detection and Management Using an Embodied Conversational Agent

Chen Wang, Beatrice Biancardi, Maurizio Mancini, Angelo Cafaro, Catherine I Pelachaud, Thierry Pun, Guillaume Chanel

Impression Detection and Management Using an Embodied Conversational Agent *

Chen Wang[1], Beatrice Biancardi[2], Maurizio Mancini[3] Angelo Cafaro[4], Catherine Pelachaud[4], Thierry Pun[1], and Guillaume Chanel[1]

[1] University of Geneva, Geneva, Switzerland
{chen.wang,thierry.pun,guillaume.chanel}@unige.ch
[2] Telecom Paris, Paris, France
beatrice.biancardi@telecom-paris.fr
[3] University College Cork, Cork, Ireland
m.mancini@cs.ucc.ie
[4] CNRS-ISIR, Sorbonne Université, Paris, France
angelo.caf@gmail.com,catherine.pelachaud@telecom.fr

Abstract. Embodied Conversational Agents (ECAs) are a promising medium for human-computer interaction, since they are capable of engaging users in real-time face-to-face interaction [1, 2]. Users' formed impressions of an ECA (e.g. favour or dislike) could be reflected behaviourally [3, 4]. These impressions may affect the interaction and could even remain afterwards [5, 7]. Thus, when we build an ECA to impress users, it is important to detect how users feel about the ECA. The impression the ECA leaves can then be adjusted by controlling its non-verbal behaviour [7]. Motivated by the role of ECAs in interpersonal interaction and the state-of-the-art on affect recognition, we investigated three research questions: 1) which modality (facial expressions, eye movements, and physiological signals) reveals most of the formed impressions; 2) whether an ECA could leave a better impression by maximizing the impression it produces; 3) whether there are differences in impression formation during human-human vs. human-agent interaction. Our results firstly showed the interest to use different modalities to detect impressions. An ANOVA test indicated that facial expressions performance outperforms the physiological modality performance (M=1.27, p=0.02). Secondly, our results presented the possibility of creating an adaptive ECA. Compared with the randomly selected ECA behaviour, participants' ratings tended to be higher in the conditions where the ECA adapted its behaviour based on the detected impressions. Thirdly, we found similar behaviour during human-human vs. human-agent interaction. People treated an ECA similarly to a human by spending more time observing the face area when forming an impression.

* Supported by the Swiss National Science Foundation under Grant Number 2000221E-164326 and by ANR IMPRESSSIONS project number ANR-15-CE23-0023

Keywords: Affective computing · Impression detection · Virtual agent · Eye gaze · Impression management · Machine learning · Reinforcement learning.

1 Introduction

Virtual agents (VAs) are widely used for human-computer interaction, as they can mimic naturalistic human communication. An Embodied Conversational Agent (ECA), one kind of VA, is able to produce and respond to verbal and nonverbal communication in face-to-face conversations [1, 2]. There are studies finding that ECAs' non-verbal behaviour is associated with emotions [3], personality traits [29] and interpersonal attitudes [4]. However, there is not much work on how ECAs' non-verbal behaviour influences formed impressions. The formed impression (e.g. favor or dislike someone) of an ECA is an internal state which may be reflected by users behaviourally [18, 20]. The formed impression could affect the interaction (e.g. willingness to interact), and the effect could even last after the interaction [6, 7]. Thus, when we build an ECA to impress users and have a good interaction, it is important to sense how users think about the VA through users' body responses. Then the impression the ECA leaves could be controlled accordingly by adapting its non-verbal behaviour. In this context, it is possible to use machine learning methods to determine the impression that an user is forming and to rely on this information to build a more engaging VA, which is able to manage the impressions they leave on users.

Impression, as an important component for social cognition and communication, has not been well explored with machine learning methods. Warmth and competence (W&C) are the most used impression dimensions in the literature about human-human and human-agent interaction [12, 18, 22, 23]. Warmth represents the intentions of the others (positive or negative), and competence stands for the consequent ability to execute those intentions. For example, if a person A meets a person B who is rude and speaks with an angry voice, A might form an impression that B is competent but rather cold. It is possible to use the signals of B to predict which impression B leaves on A and others. This is called **impression prediction** (yellow arrow in Fig.1), and most of the literature focuses on this case. On the other hand, we could use the body responses of A to detect the impression that$A$ forms of B. This is called **impression detection** (blue arrow in Fig.1) and is the main focus of this paper. The impression expressive behaviour could be conveyed through multiple modalities, including facial, gestural and physiological reactions, which may not always be congruent and have the same level of importance [5]. To the best of our knowledge, there is rarely studies with ECA which measures users' impressions and adapts its behaviour accordingly. In this paper, we would like to investigate three research questions: 1) which modality (facial, eye and physiological expressions) reveals most of the formed impressions; 2) whether an ECA could leave a better impression on users by maximizing the impression (W or C) it produces; 3) whether there are differences in impression formation during human-human

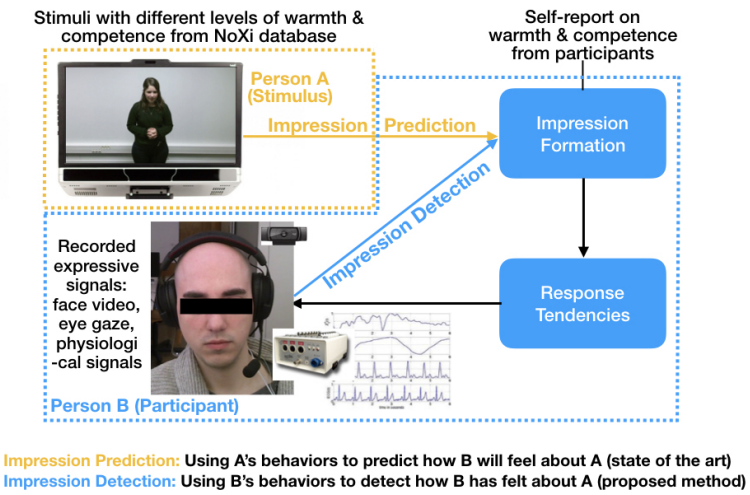


Fig. 1: Impression Formation and Detection Diagram [5]

vs. human-agent interaction. We first applied several impression detection models on each modality of an impression evoking corpus with continuous W&C self-reports. We explored the modality importance by observing the detection performance. With the learned modality importance and detection model from the first study, we built an ECA use case in which the ECA interacted with a participant and adapted its behaviour based on the detected participant impressions. We evaluated our ECA by comparing the participants' impression reports with the automatically detected impressions to investigate if our adaptive ECA could lead to changes in reported W&C. We also compared the exhibited behaviour of participants forming an impression of an ECA with the behaviour forming an impression of a human.

2 Background and Related work

2.1 Impression and Emotion recognition

Current research mainly focuses on how exhibited behaviour influences other's formation of impressions (i.e. impression prediction). For example, there are some studies on stereotype based prediction [28, 21]. Instead of studying prediction, this work focuses on detecting impressions from the expressive behaviour of the person forming the impression (blue arrow shown in Fig.1) in the W&C space, which is a new approach to impression recognition. According to [18, 25], forming an impression is associated with emotions and behaviour. To the best of our knowledge, there is only one study on assessing the formed impressions from the body signals (e.g. facial expressions and gestures) of the person forming the impression [5]. In [5], it reported that formed impressions could be detected

using multimodal signals in both multi-task and single task frameworks. However, which body signal reveals the formed impression more expressively was not discussed. Also it was not mentioned whether the detection models in [5] are suitable for real-time application such as human-agent interaction.

Studies in emotion recognition demonstrated the possibility of inferring user's emotions from multimodal signals [13]. Since emotions can be induced when forming impressions [18], this supports the possibility of assessing users' impressions from their affective expressions. Emotion recognition studies explored a variety of models using machine learning methods. These methods can be grouped in two classes based on whether temporal information is applied or not. The non-temporal models generally require contextual features. Temporal models exploit the dynamic information in the model directly. Methods such as Multilayer Perceptron (MLP), Support Vector Machine (SVM), XGBoost and Long Short Term Memory (LSTM) models are currently widely used with several topologies [13, 16, 24].

2.2 ECA Impression Management

To manipulate the impression (W&C) that the ECAs leave on users, researchers adopted findings from human-human interaction [6, 19, 27, 31] to the ECA design. Nguyen et al. [31] applied an iterative methodology that included theory from theater, animation and psychology, expert reviews, user testing and feedback, to extract a set of rules to be encoded in an ECA. To do that, they analysed gestures and gaze behaviour in videos of actors performing different degrees of W&C. In [12], it investigated the associations between non-verbal cues and W&C impressions in human-human interaction. The type of gestures, arms rest poses, head movements and smiling were annotated, as well as the perceived W&C of people who played the role of expert in a corpus of videos of dyadic natural interactions. It was found that the presence of gestures was positively associated with both W&C. A negative association was found between some arms rest poses and W&C, such as arms crossed. The smiling behaviour presented while performing a gesture could increase warmth judgements, while negatively related to competence judgements. These finds were used to guide ECA designs. Beside behaviour, the appearances of ECAs also influence the perception of W&C. For example, Bergmann et al. [11] found that human-like vs. robot-like appearance positively affected the perception of warmth, while the presence of co-speech gestures increased competence judgements.

2.3 Human-ECA Impression Formation

Since ECAs can mimic naturalistic human communication, there are studies comparing human-human interaction with human-ECA interaction. According to Wang, Joel, et al. [38], people tended to treat VAs similarly to real human beings. McRorie et al.[29] implemented four stereotypical personalities in the virtual agents. During the interaction with agents, the participants could easily identify the agents' personalities in a similar way that they identified humans

with the same personality. Anzalone et al. [8] explained the importance of assessing the non-verbal behaviour of the humans to increase the engagement during human-robot interaction. Kramer et al.[26] showed that a smiling agent did not change the inferences made by the users, but whenever the virtual agent smiled, it triggered a mimicry smile on the users. It meant that the agent succeeded in provoking a change of user behavior while not having an impact on the impression formation. Although there are studies showing that people judge or interact with ECAs similarly as with humans, it still requires investigation on whether people express their formed impression of ECAs the same way of humans.

3 Impression Detection

3.1 Impression Evoking Corpus

To build impression detection models, we relied on an impression evoked corpus reported in [5], where multimodal data of 62 participants (23 female and 39 male) was recorded while watching impression stimuli and reporting their formed impressions in W&C continuously. The data recording diagram is shown in Fig.1. The stimuli used to evoking participants' impressions are from Noxi database [14]. In each video from Noxi database, a different expert (real person) was talking about a topic of interest (e.g. cooking). The Noxi videos have corresponding continuous W&C and gesture annotations of the experts which were annotated by motivated and experienced people, with previous experience in affective annotation and background knowledge about W&C. More details of the Noxi database can be found in [14]. The original Noxi videos are too long for our experiment. Thus the stimuli used in [5] were cut and selected from the Noxi [14] database based on the warmth (range[0,1]), competence (range[0,1]) and gesture annotations (e.g.iconic). We firstly applied peak detection on the Noxi W&C annotations and selected the video clips that contain at least one change (peak) in warmth or competence. Then among the W&C changing clips, we chose the ones containing most gesture annotations. Each stimulus lasts around 2 minutes (mean = 1.92, std = 0.22) with different levels of warmth (mean = 0.56, std = 0.18) and competence (mean = 0.52, std = 0.28). The stimuli were displayed in a random sequence.

The following modalities were recorded while the participants were watching the impression stimuli: facial videos (Logitech webcam C525 & C920, sample rate 30 fps), eye movements (Tobii TX300 & T120, sample rate 300Hz and 120Hz respectively) and physiological signals (electrocardiography (ECG) and galvanic skin response(GSR), using a Biosemi amplifier, sample rate 512 Hz). At the same time participants annotated their formed impressions by pressing keyboard buttons whenever they felt a change in warmth (up & down keyboard arrow) or in competence (left & right keyboard arrow). W&C were annotated independently and could be annotated at the same time. All participants were given the same explanation about the concept of W&C before the recording. English proficiency levels were requested to be over B2 in the Common European Framework of Reference, to guarantee that the participants were able to understand and follow

experiment instructions. We used the definition of W&C in [25] and two sets of words [18] to describe W&C to help them to understand. All participants were informed the experiment content and signed a consent form. They were trained with the annotation tool and practiced before watching the stimuli. In total, the corpus contains 62 participants with 1625 minutes of multimodal recordings and W&C annotations.

3.2 Pre-processing and Feature Extraction

To prepare the recorded data for regression models, we firstly synchronized the impression annotations with multimodal recordings using the recorded triggers. The triggers are the starting timestamp of each stimulus. With the triggers and stimuli lengths, we segmented the recorded data based on each stimulus. The recorded modalities from the impression corpus have various sampling frequency ranging from 30 to 512 Hz while the impression annotations have uneven sampling frequency. We resampled the impression annotations as well as multimodal recordings or extracted features to get the same length of data. In this paper, each modality as well as annotations were resampled to 30 Hz (face video frame rate) for simplification.

To homogenize sampling frequencies of annotations and recorded signals, we used the face video frame rate as a standard and applied 1D polynomial interpolation on W&C annotations respectively to achieve the same sample rate. After the interpolation, we followed [36] and applied a 10 seconds sliding window with overlap (1 frame shift per time) to smooth warmth and competence annotations. Features were extracted from each modalities: facial video, eye gaze and physiological signals (ECG and GSR signals). We extracted the features that have been proved to work well for affective recognition [5, 13, 16, 24]. Following [13], we used action units (AU) as features which are the deconstructed representations of facial expressions [20]. The AUs were extracted on each frame using an open source tool OpenFace [10]. We had 17 AUs intensity (from 0 to 5) and 18 AUs presence (0 or 1) features. For eye movements, the 2D gaze location on the display, the 3D locations of the left and right eyes, and the gaze duration recorded by the eye tracker were taken as features. All the 9 features from eye movements are down sampled (120 Hz or 300 Hz) to the video frame rate (30 Hz). To process physiological signals, we used the TEAP toolbox [34] to extract features. We filtered out the noise with a median filter and then extracted Skin Conductance Response (SCR) from the GSR signal, heart rate (HR), heart rate variability (HRV), HR multi-scale entropy, mean heart rate over 1 minute and corresponding standard deviation from the ECG signals. We resampled the extracted features to 30 Hz instead of resampling the raw signals directly to conserve more information. All the extracted multimodal features were smoothed using the same sliding window as for annotations to get the same sample sizes. After resampling, features as well as smoothed annotations were standardized so that they all had a mean of 0 and a variance of 1 to improve gradient descent convergence and avoid having a classification bias toward high magnitude features.

3.3 Impression Detection Models

As presented in section 2.1, regression models have performed reliably in affect recognition. We tested 3 widely used regression models from different families of supervised learning algorithms: Support vector regression (SVR) from vector machines, XGBoost from ensemble methods with decision trees and Multilayer Perceptron Regression (MLP) from neural networks to detect the formed impressions in W&C dimensions. Regression models on W&C were trained and tested separately. All the aforementioned models generate predictions of a warmth (resp.competence) score at each frame (30 Hz) based on the input features. We implemented SVR and MLP using the scikit-learn library [32] and XGBoost [17] with the python XGBoost library [5]. For SVR, we used a radial basis functions kernel with gamma equals to $1/P$ as proposed in [35], where P is the number of features, and set the tolerance for the optimization stopping criterion to 1e-4. For MLP, we set 2 hidden layers with 64 neurons on each and 1 dimension output (i.e. warmth and competence detection are trained independently). We trained at most 50 epochs and applied early stopping to avoid overfitting with patience equal to 5 epochs. Mean squared error (MSE) was used as the loss function. XGBoost was set with 100 estimators and the same learning rate as MLP: 1e-3. To avoid overfitting, XGBoost and MLP were set with the same early stopping setting with a patience equal to 5.

To train and test detection models, we used a leave-one-out cross-validation scheme. We divided the data set into three partitions: 1 participant was left out for testing, the remaining data was all used for SVR training while randomly divided into two parts for MLP and XGBoost: 80 percent for training and 20 percent for validation. We rotated the left-out testing participant to estimate the model performance of all the participants. We trained and tested the 3 regression models respectively with unimodal features as well as multimodal features. We also tested multimodal detection with early fusion for combining features. That is, features from different modalities were concatenated together as the input feature matrix.

3.4 Modality Performance Analysis

We investigated the importance of each modality by calculating the Concordance Correlation Coefficient (CCC) between the detected impression and participant annotations. Significant performance differences between the modalities and regression models were tested using ANOVA.

We firstly tested significant difference in unimodal impression detection performance to check if some modalities were more accurate than others. The CCC values were shown to be normally distributed using a Shapiro test ($p = 0.31$). We thus ran a 3x3x2 between-group ANOVA, with regression model, modality, and impression dimension as factors. We did not find an effect for impression dimension (warmth or competence). A main effect of regression model was found

[5] https://github.com/dmlc/xgboost

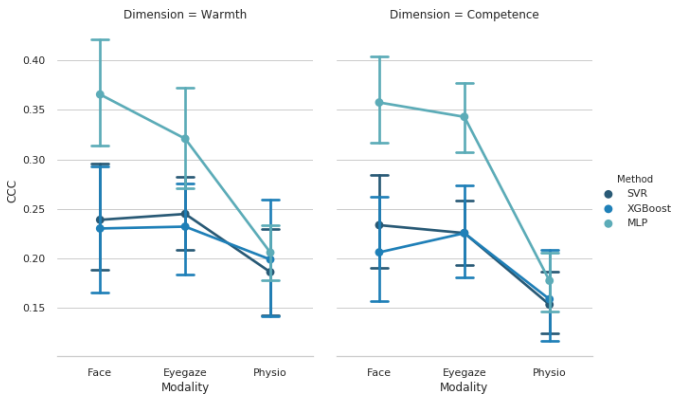


Fig. 2: Unimodal Impression Detection Performance

($F(2, 27) = 3.53, p < 0.05$). As shown in Fig.2, post-hoc tests revealed that MLP achieved higher detection accuracy than XGBoost($mean_difference = 0.19, p-adjust = 0.04$). A main effect of the modality was also found ($F(2, 27) = 4.15, p < 0.03$). Post-hoc test indicated that facial expressions performance outperformed the physiological modality performance ($mean - difference = 1.27, p - adjust = 0.02$). Although there was no significant difference between facial expressions and eye movements with all 3 regression models ($p > 0.05$), the mean CCC performance of facial modality from MLP were better than eye movements for both W&C.

We also tested the performance for multimodal impression detection. For this purpose, an early fusion strategy was employed where all modality features were concatenated in a unique feature vector. For MLP, a mean CCC of 0.652 for warmth and 0.681 for competence was obtained. This improvement of performance over unimodal detection was significant for warmth ($t = 6.63, p < 0.01$) and competence ($t = 5.71, p < 0.03$) as demonstrated by a pairwise t-test. The multimodal performance with SVR was 0.317 for warmth and 0.308 for competence. The XGBoost algorithm obtained slightly better results with a CCC of 0.332 for warmth and 0.376 for competence. These results were higher than unimodal performance but lower than multimodal MLP.

Overall, our results confirm that when individuals are unknown to us, our facial expressions reveal most the impression we've formed of the unknowns [39]. The learned modality salience could be applied in the future work of multimodal fusion at modality level for impression detection.

4 Embodied Conversational Agent Use Case

We conducted a use case in order to test our impression detection model in a user-agent real-time interaction scenario. We firstly would like to investigate whether impression detection and adaptation could improve users' formed impressions

of an ECA in real-time. Secondly we would like to compare the participants' behaviour when they are forming an impression of the ECA with the behaviour occurring in the first study, when participants observe a human stimulus. To reach those objectives, we designed an ECA which interacted with each user on a given topic. The ECA played the role of a virtual guide, introducing an exhibit about video games held at a science museum. The ECA was a black-hair female character designed based on a stereotyped-based model from [3], aiming to appear warm and competent. The ECA, named Alice, first introduced itself to the participants, and then gave them information about the exhibition. The ECA asked questions/feedback to the participant during the interaction as well (e.g. "Do you want me to tell you more about the exhibit?").

The ECA adapted its non-verbal behaviour based on the impressions detected from the users' facial expressions. The non-verbal behaviour of the ECA included gestures, arm rest poses and smiling facial expression. The behaviour was designed based on the finding from [12]. This adaptation of ECA was achieved by employing a reinforcement learning algorithm (Q-learning) that aimed at maximizing either the detected warmth or competence depending on the experimental condition. The reward of the ECA to select the most appropriate non-verbal behaviour was computed as the increase or maintenance of detected competence or warmth. The eye movements were recorded for some participants but were not used for detecting impression. Not all participants agreed to record eye movements, for example, the eye tracker cannot be used by epilepsy patients. To guarantee that we had a reliable model for impression detection, we decided to use the facial modality only. We focused on how ECA's behaviour could change the users' impressions, thus the agent appearance, voice and tone remained the same as a constant in all experimental conditions. The speech acts were scripted before the experiment.

The interaction with the ECA lasted about 3 minutes (a duration similar to the one used for the human stimulus presented in section 3.1) divided in 26 speaking turns. A speaking turn was defined as a dialog act (e.g., greeting, asking questions, describing a video game, etc.) played by the ECA and user's possible answer or verbal feedback. In the absence of user's responses (i.e. in case of user's silence lasting more than 1.5s or 4s, depending on whether the ECA just said a sentence or asked an explicit question), the ECA continued with another speaking turn. User's impression was determined using the data driven regression model presented in sections 3.3 and 4.1. The detected warmth or competence given at 30Hz were averaged over periods of 1 second without overlapping (i.e. 1 warmth or competence value per second). After each speaking turn, the the last detected warmth or competence value was sent to the reinforcement learning module to drive the ECA behaviour.

4.1 Impression Management System

The proposed system was composed of 2 main modules enabling real-time user-agent interaction as illustrated in Fig.3. The first module concerned *User's Impressions Detection* includes two sub-components: one to detect user's behaviour

(speech, facial expressions) and the second to analyse and interpret them (i.e. facial expressions were used to infer users' impressions of the ECA). The VisNet open source platform [5] was exploited to extract the user's face AUs in real-time, by running the OpenFace framework [10], and user's speech by executing the Microsoft Speech Platform[6]. Based on the extracted AUs, user's impressions were computed with the MLP model presented in section 3.3. Although the eye movements and physiological signals contributed to a better impression detection performance on average, these modalities did not increase accuracy significantly. Compared with video recordings, ECG and GSR were more invasive, and they required time and experience to attach sensors on the skin. It was not practical for our setting where participants visited a museum and barely had time for such an experiment.In the future remotely detected physiological information will be embedded into our impression detection model.

The second module was the *Agent's Impression Manager* which arbitrates verbal behaviour (i.e. what the ECA should say) and non-verbal behaviour (the ECA behaviour (e.g. smiling, gestures, etc.) accompanying speech). The ECA's speech and behaviour are dynamically selected to effectively manage impressions of W&C. The ECA impression management module was implemented with Flipper [37], a dialogue manager that, given the detected user's impressions, chooses the verbal and pre-designed non-verbal behaviour (related to W&C based on [12]) the ECA will display in the next speaking turn. The behaviour was selected according to a Reinforcement Learning (Q-learning) algorithm with the detected impressions as rewards. The reinforcement learning module defined states s (in our case these were warmth or competence level) and actions a performed by the ECA (in this paper an action is the dialogue act accompanied with verbal and non-verbal behaviour). The initial Q values ($Q(s, a)$) of actions and states were set up to 0. A reward function R was computed for each combination of state and action. In our case R was the difference between detected warmth (resp. competence) and the current warmth (resp. competence) level. The Q-learning algorithm explored all the possible next state-action pairs *(s',a')* and tried to maximize the future rewards with a discount rate γ. We maximized one dimension at a time since it is difficult to maximize both due to the halo effect [33]. The new Q values ($Q_(new)(s, a)$) are updated with the Q function. After each speaking turn, both Q table and reward table would be updated. The SAIBA-compliant AnonymAgent platform supported the generation of behaviour and computed the corresponding animation of the ECA [5]. More details on the interactive system can be found in [5].

To evaluate our impression detection model performance, we set 3 conditions for the ECA: Warmth, Competence and Random. Under Warmth and Competence conditions, the ECA performed the behaviour that the Impression Manager chose. That is, during the experiment, the ECA performed one of the pre-defined gestures according to the Q-learning method and in order to maximize either warmth or competence. Under the Random condition, the ECA performed behaviour that was randomly selected among the set of possible behaviour.

[6] https://www.microsoft.com/en-us/download/details.aspx?id=27225

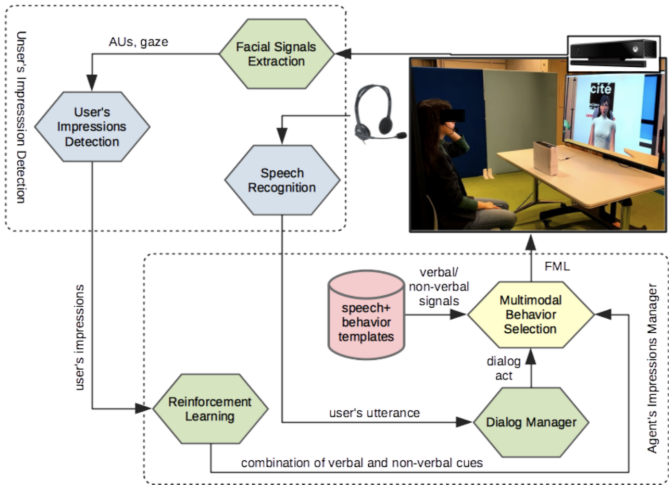


Fig. 3: The Impression Management System [5]

4.2 Collected Data

All participants were visitors at the museum who were voluntarily participating. They all signed a consent form before the recording. In total, we collected data from 71 participants who were randomly assigned to each condition. We got 25 participants in the Warmth condition, 27 in the Competence condition and 19 in the Random condition. Upper body videos (Kinect V2, sample rate 30 fps) were recorded for all participants and we also collected eye movements (Tobii Pro Nano, sample rate 60 Hz) from 19 participants (8 for Warmth, 8 for Competence and 3 for Random). Participants answered a questionnaire after the interaction with the ECA to report their overall formed impressions of the ECA in the W&C dimension (4 items concerning warmth, 4 concerning competence with a 7 point Likert scale, according to [9]). In order to group together the 4 reported values for warmth and the 4 for competence, Cronbach's alphas were computed on the scores. Good reliability was found for both W&C with $\alpha = 0.85$ and $\alpha = 0.81$ respectively. The mean of these items were calculated separately in order to have one warmth score and one competence score for each participant.

5 Result and Discussion

5.1 Impression management efficiency

To evaluate our real-time impression model performance, we firstly compared the detected warmth or competence from the facial modality with the reported impressions from the questionnaire. Secondly, we checked whether the trends of the detected warmth or competence were increasing in their respective condition. In other words we verified that warmth (resp. competence) was overall increasing in the warmth (resp. competence) condition.

The detected W&C given at 30Hz were averaged over periods of 1 second without overlapping (i.e. one warmth or competence value per second). Although only the last detected warmth or competence value after each speaking turn was sent to drive the ECA behaviour, we recorded all the detected warmth or competence values. We used the mean value of the detected impression over the whole interaction period and call it the *mean impression* of the participant. We also calculated the average value of the last 10 seconds of the detected impression scores as the *late detected impression*. We standardized detected W&C values (both average & late average) and self-reported ones to remove the scale influence for CCC. The CCC between reported impressions and late detected impression ($W = 0.38, C = 0.42$) were higher than those between average impression in both warmth (0.29) and competence (0.31) dimensions, which means the late detected impression is closer to the self-reported impression. To test differences in participants reported impressions in the three difference conditions (Random, Competence adaptation and Warmth adaptation), a one-way ANOVA was employed. The results showed that participants in the Competence condition gave higher scores than participants in the Random conditions($F(2, 32) = 3.12, p < 0.05$). There was no significant effect between Random condition and Warmth conditions, though the mean impression scores were higher than the Random condition [5].

The results showed that participants' impressions could change during the interaction. The later detected impression was closer to the participants' reported impression. Participants' ratings tended to be higher in the W&C conditions in which the ECA adapted its behaviour based on detected impressions, compared to the Random condition. In particular, the results indicated that we managed to manipulate the impression of competence with our adaptive ECA.

Under Warmth and Competence conditions, the ECA changed its behaviour in order to maximize participants' perception of warmth and competence. Thus we calculated the global deterministic trend of detected warmth/competence to check whether they were increasing consistently. The trend was determined by computing the linear regression coefficient of the detected impressions using the python StatsModel module. Under the Warmth condition, our ECA managed to increase warmth or keep warmth in a high level for the majority of participants (15 out of 25). In the Competence condition, competence was increasing for only 13 out 27 participants. This could be caused by inaccuracy of the detected impression or the agent impression management module (e.g. choose an ECA behaviour which is supposed to increase competence but actually causes competence decrements).

5.2 Impression Formation of Humans and ECA

To compare how people behave when forming an impression of a person and an ECA, we analyzed gaze patterns during these two type of interactions. For the human-human interaction, we extracted patterns of participants from the first study when they were watching the human stimuli presented in Section 3.1. This allowed us to study the differences in behaviour when forming an impression of

a human and an agent. We firstly rejected all samples without gaze detection (because of blinking or participant eye drifting). We then extracted the face area of our human stimuli and ECA using 67 landmarks extracted by OpenFace[10]. The face area was defined as the smallest rectangle area (green rectangle of Fig 4) containing all facial landmarks. If the gaze locates within the rectangle area, we assume that the participant is looking at the face area. If the gaze locates out of the rectangle, we assume that the participant is looking at other regions. We also used the line connecting landmark 8 and landmark 27 to separate the left from the right hemiface shown in Fig.4. For human stimuli, we counted the

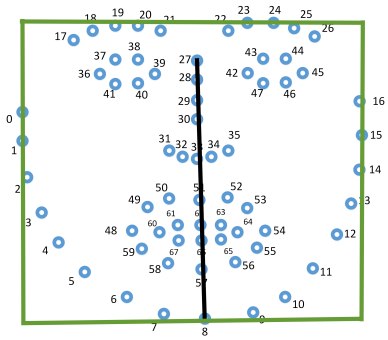


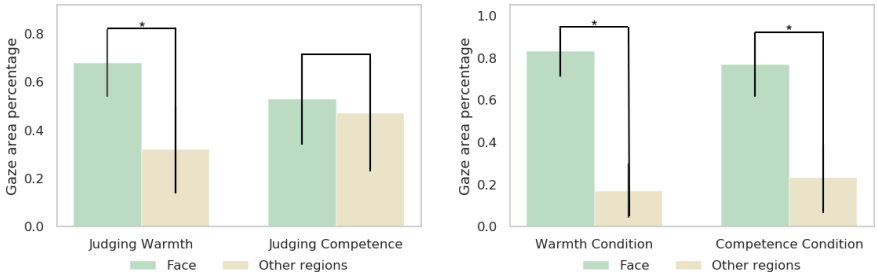
Fig. 4: Landmarks from OpenFace [10]. The green rectangle defines the face area. The vertical black bar separates the left from the right hemiface.

percentage of gazes located within the face area when the participants reported impression changes. For this purpose, we extracted a 2 seconds window centered around each W&C annotations. For the ECA, we did a similar processing as human stimuli, however, since in this case we do not have annotations all along the interaction, we took the whole interaction under Warmth and Competence conditions separately to compute the percentage of gazes on the face area.

We tested if participants were looking more at the face than at other regions using a Chi-square test. As shown in Fig. 5a, participants spent significantly more time gazing at the face area of the human stimulus when judging warmth ($p(5.51) = 0.041$). For competence, no significant difference was found ($p > 0.05$) and it appeared that participants spent similar amount of time looking at the face and the other regions. Although the setting for human-human interaction and human-ECA interaction was not exactly the same, people showed similar eye behaviour when interacting with the ECA (shown in Fig.5b) by spending significantly more time looking at the face area ($p < 0.03$), but this time for both the Warmth and Competence conditions. To compare eye behaviour between the human stimuli and human-ECA interactions, we ran a 2x2 Chi-squared test with the experiment (human vs. ECA) and the impression dimensions (warmth vs. competence) as independent variables. The result of this test was not significant

($p = 0.94$) indicating that participants' eye behaviour was similar under all conditions.

For both human-human interaction and human-virtual agent interaction, face modality played an important role in forming impression. When interacting with an ECA, people mainly focused on the face area. While watching human stimuli, people also spent time to glance at the background, stimuli gestures, clothes and so on. This confirmed the finding of Cassell et al. [15], the modeling of the ECA face is an important component for the impact on the user. .



(a) Participants' gaze area when they reported changes in W&C space while watching human stimuli (* $p < 0.05$)

(b) Participants' gaze area under W&C conditions while interacting with ECA (* $p < 0.05$)

Fig. 5: Gaze area in human-human vs. human-agent interaction

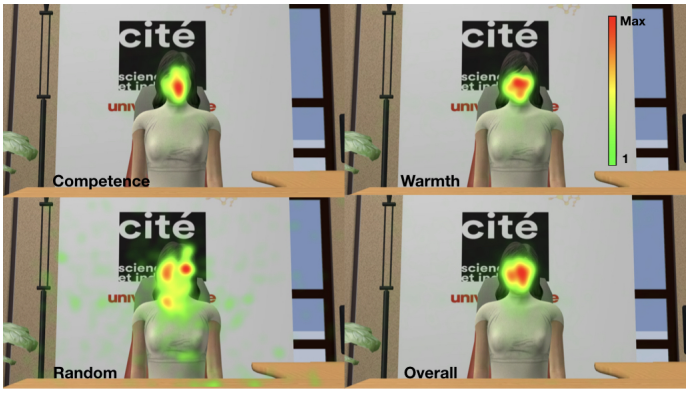
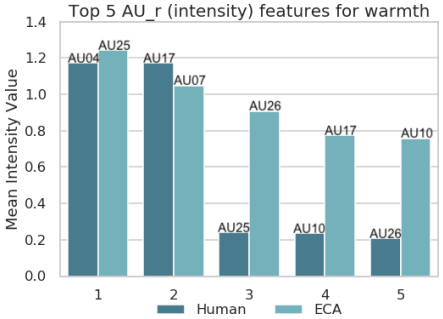
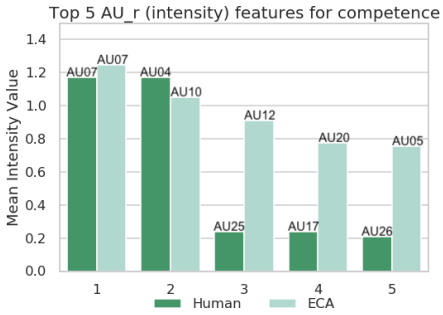


Fig. 6: Participants' gaze area while interacting with ECA in different conditions

According to [30], people demonstrate significant left-sided facial asymmetry when expressing emotions (i.e. facial expressions are more intense and faster on the left side). In addition people are more sensitive to the left hemiface of



(a) Top 5 presented facial features in when judging warmth



(b) Top 5 presented facial features when judging competence

Fig. 7: Presented facial features when form an impression

others for emotion perception. In our study, this effect was found both when interacting with human stimuli and ECA. For human stimuli, participants spent significantly more time looking at the left hemiface both for judging W&C with Chi-square test $chi - square = 6.27/6.86, p < 0.05$. But there was no significant difference between judging W&C. That is, when people looked at the face area for impression judgement, they looked more at the left hemiface no matter judging warmth or competence. While interacting with the ECA (Fig.6), participants paid more attention on the left side of the agent face in all three conditions (Warmth, Competence and Random). There is no significant difference from interacting with humans with ANOVA test $p > 0.05$. Within the three conditions, there were slight differences of the eye behaviour, for example, under the Random condition, the eye gaze was less clustered compared with the other two conditions. This might be caused by the small amount of data in the Random condition (3 participants with eye movement recording).

Beside eye movements, we also analyzed participants' facial expressions when they interacted with the ECA and human stimuli. Similar to the gaze area analysis, we used the intensity values of 17 AUs through the whole ECA interaction under Warmth and Competence conditions. For human-human interaction, we took the 2-second windows centered at W&C annotations respectively. The AU intensity (1 value per frame ranging from 0 to 5 for 1 AU) presents how intense the detected AU is. We calculated the mean intensity of all 17 AUs and selected the top five AUs for W&C separately. It was found that people showed different AUs more often when evaluating other's warmth or competence. However, when judging warmth, there are 3 common AUs among the top 5 that appeared on participants' faces for both human-human and human-ECA interaction as shown in Fig.7a. That was AU25 *lips part*, AU07 *lid tightener* and AU10 *upper lip raiser*. Although they had different ranking with ECA and human, they all presented intensively when participants were processing warmth related information. While for judging competence (Fig.7b), there was only one mutual AU that revealed intensively under both human-human and human-ECA interaction, which was

AU07 *lid tightener*. This was the most intense AU that appeared on participants for judging competence. This AU also presented when assessing warmth. Another interesting finding was that the mean AU intensity of the 17 AUs was higher ($mean - diff - warmth = 0.118, mean - diff - comp = 0.335$) when participants was interacting with the ECA other than human, no matter judging warmth nor competence. That means participants were more expressive when facing an ECA than a human stimulus.

6 Conclusion and Future Work

Our results showed the interest to use different modalities to detect formed impressions, namely facial expressions, eye movements and physiological reactions. Among all modalities, facial expressions achieve the highest accuracy with the MLP model. Secondly, our results presented the possibility of creating an adaptive ECA by detecting users' impressions from facial expressions. In the ECA use case, our results showed the consistency in late detected impression scores from facial expressions and participants' self-reports. Participants' ratings tended to be higher in the conditions in which the ECA adapted its behaviour based on the detected impressions, compared with the randomly selected ECA behaviour. Thirdly, we found similar behaviour in impression formation during human-human vs. human-agent interaction. People treated the ECA similarly as humans by spending significantly more time observing the face area when forming an impression. That indicated that participants' impressions could be manipulated by using non-verbal behaviour, particularly facial expressions and possibly gestures. Participants also presented similar facial expressions when they formed an impression of an ECA or a human, while they facially expressed more when they faced an ECA. These insights could be used to better understand the theoretical basis for impression formation and could be applied in creating adaptive ECAs.

Our work has its limitations and many aspects remain to be explored in detecting and managing impressions for ECA. Our work targets on the non-verbal behaviour of an ECA. According to [18, 22, 25], appearance (e.g. physical aspect and clothing style) could also influence the impression formation. In our case, the ECA did not change its appearance and we regarded it as an constant. However, with different appearance setting, it may enhance or decrease the impression perception caused by non-verbal behaviour. For impression detection, more multimodal fusion methods will be explored to improve the detection performance. The different facial expressive behaviour while facing an ECA than a human, indicates training machine learning model with human-ECA interaction data to improve the detection accuracy. With a better detection model, the ECA could more adequately choose the correct behaviour. Besides, there may be better solutions than reinforcement learning for selecting impression evoking behaviour to improve the ECA performance.

References

1. Angelo Cafaro, Hannes Hogni Vilhjalmsson, and Timothy Bickmore. 2016. First Impressions in Human–Agent Virtual Encounters. ACM Trans. Comput.-Hum. Interact. 23, 4, Article 24 (Aug. 2016), 40 pages.
2. Mark Ter Maat, Khiet P Truong, and Dirk Heylen. 2010. How Turn-Taking Strategies In uence Users' Impressions of an Agent. In IVA, Vol. 6356. Springer, 441–453
3. Catherine Pelachaud. 2009. Modelling multimodal expression of emotion in a virtual agent. Philosophical Transactions of the Royal Society of London B: Biological Sciences 364, 1535 (2009), 3539–3548.
4. Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. 2013. From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In International Workshop on Intelligent Virtual Agents. Springer, 263–274.
5. Wang, Chen, Thierry Pun, and Guillaume Chanel. "Your Body Reveals Your Impressions about Others: A Study on Multimodal Impression Detection." 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE, 2019.
6. Biancardi, Beatrice, et al. "A Computational Model for Managing Impressions of an Embodied Conversational Agent in Real-Time." 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2019.
7. Goffman, Erving. The presentation of self in everyday life. London: Harmondsworth, 1978.
8. Salvatore M Anzalone, So ane Boucenna, Serena Ivaldi, and Mohamed Chetouani. 2015. Evaluating the engagement with social robots. International Journal of Social Robotics 7, 4 (2015), 465–478
9. Juan I Aragones, Lucia Poggio, Veronica Sevillano, Raquel Perez-Lopez, and Maria-Luisa Sanchez-Bernardos. 2015. Measuring warmth and competence at inter-group, interpersonal and individual levels/Medicion de la cordialidad y la competencia en los niveles intergrupal, interindividual e individual. Revista de Psicologia Social 30, 3 (2015), 407–438.
10. T. Baltruvsaitis, P. Robinson, and L.-P. Morency. 2016. Openface: an open source facial behavior analysis toolkit. In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 1–10.
11. Kirsten Bergmann, Friederike Eyssel, and Stefan Kopp. 2012. A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In International Conference on Intelligent Virtual Agents. Springer, 126–138.
12. Beatrice Biancardi, Angelo Cafaro, and Catherine Pelachaud. 2017. Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions. In Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 341–349.
13. Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Camp- bell, Charlie Dagli, and Thomas S Huang. 2016. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 97–104.
14. Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes 871 Torres Torres, Catherine Pelachaud, Elisabeth Andre, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In

Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 350–359.
15. Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. Embodied conversational agents. MIT press.
16. Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal multi-task learning for dimensional and continuous emotion recognition. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 19–26.
17. Tianqi Chen and Carlos Guestrin.2016.Xgboost:A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge 879 discovery and data mining. ACM, 785–794
18. Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. Advances in experimental social psychology 40 (2008), 61–149.
19. Duchenne, d B, The mechanism of human facial expression or an electro-physiological analysis of the expression of the emotions (A. Cuthbertson, Trans.), New York: Cam-bridge University Press.(Original work pub-lished 1862), 1990
20. Paul Ekman and Dacher Keltner. 1997. Universal facial expressions of emotion. Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture (1997), 27–46.
21. Golnoosh Farnadi, Shanu Sushmita, Geetha Sitaraman, Nhat Ton, Martine De Cock, and Sergio Davalos. 2014. A multivariate regression approach to personality impression recognition of vloggers. In Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition. ACM, 1–6.
22. Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. Trends in cognitive sciences 11, 2 (2007), 77–83.
23. Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of personality and social psychology 82, 6 (2002)
24. Hatice Gunes and Maja Pantic. 2010. Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions (IJSE) 1, 1 (2010), 68–99.
25. Charles M Judd, Laurie James-Hawkins, Vincent Yzerbyt, and Yoshihisa Kashima. 2005. Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. Journal of personality and social psychology 89, 6 (2005)
26. Nicole Krämer, Stefan Kopp, Christian Becker-Asano, and Nicole Sommer. 2013. Smile and the world will smile with you—The effects of a virtual agent's smile on users' evaluation and behavior. International Journal of Human-Computer Studies 71, 3 (2013), 335–349
27. Fridanna Maricchiolo,Augusto Gnisci,Marino Bonaiuto,and Gianluca Ficca.2009. Effects of different types of hand gestures in persuasive speech on receivers' evaluations. Language and Cognitive Processes 24, 2 (2009), 239–266.
28. Mel McCurrie, Fernando Beletti, Lucas Parzianello, Allen Westendorp, Samuel Anthony, and Walter J Scheirer. 2017. Predicting first impressions with deep learning. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 518–525.
29. Margaret McRorie, Ian Sneddon, Etienne de Sevin, Elisabetta Bevacqua, and Catherine Pelachaud. 2009. A model of personality and emotional traits. In International Workshop on Intelligent Virtual Agents. Springer, 27–33.

30. Caridad R Moreno, Joan C Borod, Joan Welkowitz, and Murray Alpert. 1990. Lateralization for the expression and perception of facial emotion as a function of age. Neuropsychologia 28, 2 (1990), 199–209.
31. Truong-Huy D Nguyen, Elin Carstensdottir, Nhi Ngo, Magy Seif El-Nasr, Matt Gray, Derek Isaacowitz, and David Desteno. 2015. Modeling Warmth and Competence in Virtual Characters. In International Conference on Intelligent Virtual Agents. Springer, 167–180.
32. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830.
33. Seymour Rosenberg, Carnot Nelson, and PS Vivekananthan. 1968. A multidimensional approach to the structure of personality impressions. Journal of personality and social psychology 9, 4 (1968), 283
34. Mohammad Soleymani, Frank Villaro-Dixon, Thierry Pun, and Guillaume Chanel. 2017. Toolbox for Emotional feAture extraction from Physiological signals (TEAP). Frontiers in ICT 4 (2017), 1.
35. Johan AK Suykens. 2001. Nonlinear modelling and support vector machines. In IMTC 2001. proceedings of the 18th IEEE instrumentation and measurement technology conference. Rediscovering measurement in the age of informatics (Cat. No. 01CH 37188), Vol. 1. IEEE, 287–294.
36. Nattapong Thammasan, Ken-ichi Fukui, and Masayuki Numao. 2016. An investigation of annotation smoothing for eeg-based continuous music-emotion recognition. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 003323–003328.
37. Jelte van Waterschoot, Merijn Bruijnes, Jan Flokstra, Dennis Reidsma, Daniel Davison, Mariet Theune, and Dirk Heylen. 2018. Flipper 2.0: A Pragmatic Dialogue Engine for Embodied Conversational Agents. In Proceedings of the 18th International Conference on Intelligent Virtual Agents. ACM, 43–50.
38. Yuqiong Wang, Joe Geigel, and Andrew Herbert. 2013. Reading personality: Avatar vs. human faces. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. IEEE, 479–484.
39. Megan L Willis, Romina Palermo, and Darren Burke. 2011. Social judgments are influenced by both facial expression and direction of eye gaze. Social cognition 29, 4 (2011), 415–429.