Abstract

How to build virtual agents that establish rapport with humans? According to Tickle-Degnen and Rosenthal [4], the three essential components of rapport are mutual attentiveness, positivity and coordination. In our previous work, we designed an embodied virtual agent to establish rapport with a human speaker by providing rapid and contingent nonverbal feedback [13] [22]. How do we know that a human speaker is feeling a sense of rapport? In this paper, we focus on the positivity component of rapport by investigating the relationship of human speakers’ facial expressions on the establishment of rapport. We used an automatic facial expression coding tool called CERT to analyze the human dyad interactions and human-virtual human interactions. Results show that recognizing positive facial displays alone may be insufficient and that recognized negative facial displays was more diagnostic in assessing the level of rapport between participants.

1. Introduction

Rapport, as Merriam-Webber dictionary defines it, is a relation marked by harmony, conformity, accord and affinity. When you engage in a good conversation with someone, that feeling of flow and connection is what formally known as rapport. But what makes you feel that you have good rapport with someone, including a virtual agent? Is it what you said? Is it the smile? How can the virtual agent know that you are feeling rapport with him or her?

Research shows that nonverbal signals are more indicative of rapport than verbal signals in dyad interaction [5]. Bernieri et al. argues that gesturing, interactional synchrony (i.e., coordination), and proximity were particularly potent indicators of rapport [10]. Tickle-Degnen and Rosenthal equate rapport with behaviors indicating positive emotions (e.g. head nods or smiles), mutual attentiveness (e.g. mutual gaze), and coordination (e.g. postural mimicry or synchronized movements) [4]. In our previous work, we explored the potential of nonverbal behavior synchrony by building embodied agents to establish rapport with humans [13, 22]. By providing contingent feedback (e.g., nods) to vocal or behavioral cues of a human speaker, the virtual agent listener created strong feelings of rapport and increased engagement and speech fluency. Several other research groups are also exploring the potential of embodied agents to establish rapport with humans through similar contingent nonverbal behavior [16-21].

However, one of the limitations of our previous embodied agents work is that the behavior contingency was limited to head nods and body posture mimicking. As Tickle-Degnan and Rosenthal pointed out, positive emotions are part of the fundamental nonverbal behavior structure of rapport. Although positive emotions can be conveyed by head nods, facial expressions, such as smile, are undoubtedly the most universal and powerful expressions of positive emotions. Previous research investigating how people use nonverbal signals to judge rapport identified smile as one of the criteria people often use along with behavior synchrony [5].

In this paper, we investigate the role facial expressions plays in the building of rapport. The goal of the investigation is to computationally model speakers’ behaviors to inform the virtual humans about the level of rapport their human counterparts feel. Although behavior contingency, such as synchrony of facial expressions, is important in building rapport, the focus of this paper is to investigate whether facial expressions can be indicators of rapport. Our analysis is conducted based on both human-human interaction data and human-virtual-human interaction data. The virtual human is the virtual Rapport Agent we previously developed [13, 22]. It will be explained in detail in the next section.

2. Virtual Rapport Agent

Inspired by findings that feelings of rapport are correlated with simple contingent behaviors between speaker and listener, including behavioral mimicry [11] and back-channeling (e.g., nods [12]), we designed a virtual human called the virtual Rapport Agent [13] to elicit the harmony, fluidity, synchrony, and flow one feels when achieving rapport. The Rapport Agent was designed to work in a particular “face-to-face monologs” paradigm where a human participant tells a story to a silent but attentive listener. In such settings, human listeners can indicate rapport through a variety of nonverbal signals (e.g., nodding, smiling, postural mirroring, etc.) The Rapport Agent attempts to replicate these behaviors through a real-time analysis of the
speaker’s voice, head motion, and body posture, providing rapid nonverbal feedback. Although as previously indicated, the contingent feedback provided by the Rapport Agent was currently limited to nodding and posture mirroring. The Rapport Agent uses a vision based tracking system and signal processing of the speech signal to detect features of the speaker and then uses a set of reactive rules to drive the listening mapping.

Studies evaluating the contingent feedback of the virtual Rapport Agent showed that people who interacted with the Rapport Agent felt stronger feelings of rapport, increased engagement and improved speech fluency compared to people who interacted with agents that provided non-contingent feedback [13]. Other studies also show that people high in trait-anxiety are more engaged speaking with the Rapport Agent than they are speaking with a stranger face-to-face [14]. Latest study showed that people who are more agreeable established more rapport with the Rapport Agent and suffered less speech disfluency [15]. These studies demonstrated that the sense of rapport can be experimentally manipulated in a lab environment and the Rapport Agent can be a useful tool to study behaviors associated with rapport.

3. Facial Action Coding System

To analyze the facial expressions, we used the Facial Action Coding System (FACS) [7]. The FACS is arguably the most widely used method for coding facial expressions in the behavioral sciences. The system describes facial expressions in terms of 46 component movements, which roughly correspond to the individual facial muscle movements. FACS provides an objective and comprehensive way to analyze expressions into elementary components. Because it is comprehensive, FACS has proven useful for discovering facial movements that are indicative of cognitive and affective states [8].

In the investigator’s guide to FACS [9], Ekman and Friesen described the action units that are generally associated with facial expressions of different emotions. For example, facial expressions of joy typically include the activation of AU 12 (Lip Corner Puller) and AU 6 (Cheek Raise). AU 9 (Nose Wrinkle) or AU 10 (Upper Lip Raise) is often seen in facial expression of disgust. Following the investigator’s guide and given the action units we can automatically code (described below), we used AU 12 as an indication of positive emotional facial expressions and AU 4 (Brow Lower), AU 9 and AU 10 as indication of negative emotional facial expressions (Figure 1). AU 6 is not included in the analysis since current version of CERT does not code AU 6. Positive and negative emotional facial expressions can certainly include other action units. However, from the actions units that can be automatically detected by CERT so far, these are the most commonly associated with positive and negative emotional facial expression.

4. CERT

The primary limitation to the widespread use of FACS is the time required to code. FACS was developed for coding by hand, using human experts. It takes over 100 hours of training to become proficient in FACS, and it takes approximately 2 hours for human experts to code each minute of video.

Table 1: Action Units automatically coded by CERT. The AUs highlighted are the ones used as indicators of positive and negative emotional facial expression in this paper.

<table>
<thead>
<tr>
<th>Action Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Inner Brow Raise</td>
</tr>
<tr>
<td>2</td>
<td>Outer Brow Raise</td>
</tr>
<tr>
<td>4</td>
<td>Brow Lower</td>
</tr>
<tr>
<td>5</td>
<td>Upper Lid Raise</td>
</tr>
<tr>
<td>9</td>
<td>Nose Wrinkle</td>
</tr>
<tr>
<td>10</td>
<td>Upper Lip Raise</td>
</tr>
<tr>
<td>12</td>
<td>Lip Corner Puller</td>
</tr>
<tr>
<td>14</td>
<td>Dimpler</td>
</tr>
<tr>
<td>15</td>
<td>Lip Corner Depresser</td>
</tr>
<tr>
<td>17</td>
<td>Chin Raiser</td>
</tr>
<tr>
<td>20</td>
<td>Lip Stretch</td>
</tr>
</tbody>
</table>

To analyze the facial expression more efficiently, we processed our video data through the Computer Expression Recognition Toolbox (CERT) developed by University of California at San Diego [1]. CERT is a user independent fully automatic system for real time recognition of facial actions from the Facial Action Coding System (FACS). The current version of CERT produces an 11 channel output stream. Each output stream channel consists of one real valued number for an AU, for each frame of the video. The real valued number indicating the distance to the separating hyper-plane for each classifier Support Vector Machine (SVM). Previous work showed that the distance to the separating hyper-plane (the margin) contained information about action unit intensity [6]. The 11 channels (AUs) CERT outputs are shown in Table 1. Previous work [3] shows that CERT performs comparably to human observers in the discrimination of distinct basic emotion classes and judgments of the similarity between distinct basic emotions.
5. Analysis of Human-Human Interaction

5.1. Data Description

To study correlations between rapport and facial expressions, we first conducted analysis on data collected from a human face-to-face interaction study.

In this study, we recruited 66 people (62% women, 38% men) from the general Los Angeles area. They were recruited by responding to recruitment posters posted on Craigslist.com and were compensated $20 for one hour of their participation. On average, the participants were 36.4 years old (min = 21, max = 60, std = 10.0) with 15.6 years of education (min = 11, max = 20, std = 1.9).

Participants came to the lab in pairs and were randomly assigned the role of speaker and listener. During the experiment, the speaker viewed a short segment of a video clip taken from the Edge Training Systems, Inc. Sexual Harassment Awareness video. The video clip was merged from two clips: The first, “CyberStalker,” is about a woman at work who receives unwanted instant messages from a colleague at work, and the second, “That’s an Order!,” is about a man at work who is confronted by a female business associate, who asks him for a foot massage in return for her business.

After viewing the video, the speaker retold the stories portrayed in the clips to the listener. The speaker and the listener sat approximately 8 feet apart from each other. In the end, both participants completed a post-questionnaire. During the interaction, the participant was videotaped.

We constructed a 10-item rapport scale (coefficient alpha = .89), presented to speakers and the listener in the post-questionnaire. This scale was measured with an 8 point metric (1 = Disagree Strongly; 8 = Agree Strongly). The average of the speaker self-reported rapport is 5.29 and the average for the listener is 5.00.

5.2. Result

Data from 1 participant was excluded due to missing data. As a result, data from 32 sessions were included in the analysis.

To process the CERT output, we adopted the statistical method Littlewort and her colleagues used to differentiate posed and genuine pain [2]. We calculated mean of Z-scores for each participant (speaker only) and each AU detector as $Z = (x - \mu) / \sigma$, where $(\mu, \sigma)$ are the mean and variance for the output of the first 120 frames of each participant’s video. During the first 4 seconds of the recording (first 120 frames), participants maintained a relatively neutral face. Thus videos from the first 4 seconds were used as a baseline for each participant.

To study whether positive and negative facial expressions are indications of rapport, we conducted a stepwise linear regression using rapport as dependent variable and AU 4, AU 9, AU 10 and AU 12 as independent variable. The model kept AU 10 and excluded AU 4, AU 9 and AU 12. The resulting model with AU 10 is statistically significant ($F=5.67, p=.025, \beta=-.43$). This indicates that negative facial expression, such as disgust (display of AU 10) in the speaker is a significant predictor of lack of rapport.

6. Analysis of Human-Virtual-Human Interaction

To test whether the results from human-human interaction study can be replicated in the human-virtual-human study, we conducted further analysis on the data collected from a virtual Rapport Agent study. In the human face-to-face study, the average level of rapport is relatively high. Using a virtual human, we can control behaviors agent exhibit to elicit different levels of feelings rapport, for example the agent can show contingent or non-contingent feedback, and display or not display facial expressions. Previous studies show virtual agent that exhibit proper contingent nonverbal behavior can induce as high as the levels of rapport established when one interacts with a real human listener [13].

6.1. Data Description

In this study, we recruited 144 people (62.5% women, 37.5% men) from the general Los Angeles area. They were recruited by responding to recruitment posters posted on Craigslist.com and were compensated $20 for one hour of their participation. On average, the participants were 39.5 years old (min = 19, max = 60, std = 11.6) with 15.8 years of education (min = 12, max = 20, std = 1.6).

Participants were divided into three groups. Each group interacted with one of the three virtual agents. The first virtual agent is a “good virtual listener” (the “Responsive” condition). The agent continuously gazes at the speaker and exhibits attentive listening behaviors (e.g. head nods and posture shifts) that have previously been demonstrated to create self-reported feelings of rapport [13]. The second virtual agent, a “not responsive listener” (the “Non-responsive” condition), gazes continuously at the speaker, but does not provide attentive listening feedback (it does exhibit random idle-time behaviors such as blinking). Finally, the “ignoring listener” (the “Ignore” condition), does not maintain gaze with the speaker (it gazes randomly about the room) and does not provide attentive listening feedback.

The study design was a between-subjects experiment with three conditions: Responsive (n = 51), Non-responsive (n = 47), and Ignore (n = 46), to which participants were randomly assigned.

During the experiment, the participant first viewed one of two videos. One of the videos was a Tweety and Sylvester cartoon. The other video is the “CyberStalker” clip taken from the Edge Training Systems, Inc. Sexual Harassment Awareness video.

After viewing the video, the participant retold the
stories portrayed in the clips to a human listener, who is a confederate in this study. The participant sat in front of a computer monitor and sat approximately 8 feet apart from the listener, who sat in front of a TV. They could not see each other directly, being separated by a screen. The participant saw the virtual agent displayed on the computer monitor and was told that the virtual agent represents the human listener. While the participant spoke, the listener could see a real time video image of the participant on the TV. Next, the participant completed a questionnaire about the contents of the video he/she just saw. Later, the participant watched the remaining of the two videos and retold the stories portrayed in the clips to the listener. After that, the participant completed the post-questionnaire. No participants indicated that they believed the listener was a confederate in the study. During the interaction, the participant was videotaped.

We used the same post-questionnaire on rapport used for the human face-to-face study. An ANOVA test comparing the self-report of rapport between participants interacted with different virtual agents confirms that participants felt significantly different levels of rapport when interactions with different agents ($M_{\text{Responsive}}=4.56$, $M_{\text{Non-responsive}}=3.60$, $M_{\text{Ignore}}=3.35$, $F=10.79$, $p<.001$). However, there was no significant difference on self-reported rapport between participants interacted with the Responsive agent and participants in the human face-to-face study ($p=.146$). This is consistent with findings from previous studies [13].

6.2. Result

Data from 12 participants were excluded due to missing data and technical difficulties during the experiment. Further, data from another 13 participants were excluded because CERT had difficulty tracking participants’ face in the video due to lighting and camera angle issue. Since accuracy of the facial expression analysis relying on locating the face first, results from these 13 participants were excluded. As a result, data from 119 sessions were included in the analysis, 38 in the Responsive condition, 41 in the Non-responsive condition and 40 in the Ignore condition.

Unlike the participants in the human face-to-face study, who viewed and discussed two Sexual Harassment Awareness video, participants in the Virtual Rapport Agent study viewed and discussed one Sexual Harassment Awareness video and one Tweety and Sylvester cartoon. These two videos are of different emotional valance: the Sexual Harassment Awareness video is more of negative emotional valance while the Tweety and Sylvester cartoon is more of positive emotional valance. Before comparing result to the human face-to-face study, we’d first like to see whether there is any differences in the facial expressions display when discuss these two videos, in particular whether there are more positive facial expressions when participants discussed the positive emotional video. Paired-sample t-test shows that there was significantly more activity of AU 12 when participants were discussing the Tweety & Sylvester cartoon than when they were discussing the Sexual Harassment Awareness video ($p=.008$). This means that participants smiled more when discussing the more positive emotional topic.

To study whether positive and negative facial expressions are indications of rapport, we conducted stepwise liner regression using rapport as dependent variable and AU 4, AU 9, AU 10 and AU 12 as independent variable. The model kept AU 9 and excluded AU 4, 10 and 12. The resulting model with AU 9 is statistically significant ($F=6.53$, $p=.012$, Beta=-.234). This confirms the results from the human face-to-face study that that negative emotional facial expression, such as disgust (display of AU 9), in the speaker is a significant predictor of lack of rapport.

Table 2: Comparison of Stepwise Linear Regression results between Human-Human interaction study and Human-Virtual-Human interaction study. The AU in this table indicates the AU that’s the significant predictor of rapport.

<table>
<thead>
<tr>
<th>Study</th>
<th>AU</th>
<th>$F$</th>
<th>$p$</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Human</td>
<td>10</td>
<td>5.67</td>
<td>.025</td>
<td>-.430</td>
</tr>
<tr>
<td>Human-Virtual-Human</td>
<td>Overall</td>
<td>9</td>
<td>6.53</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>SH video</td>
<td>9</td>
<td>8.50</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>1st video</td>
<td>9</td>
<td>11.92</td>
<td>.001</td>
</tr>
</tbody>
</table>

For a more direct comparison with the human face-to-face study, we further conducted stepwise liner regression using rapport as dependent variable and AU 4, AU 9, AU 10 and AU 12 displayed when discussing the Sexual Harassment Awareness video as independent variable. Again, the model kept AU 9 and excluded AU 4, 10 and 12. The resulting model with AU 9 is statistically significant ($F=8.50$, $p=.004$, Beta=-.265). Another difference between the human face-to-face study and this study is that the virtual agents the participants interacted with didn’t display any facial expressions. It is possible that participants “synchronized” their facial expressions with the virtual agents as the interaction moved along. Since previous studies suggest that accurate judgment of rapport can be made by observing non-verbal signals presented in a thin slice of an interaction instead of the entire interaction [23][10][24], we sliced each participant’s videos into two segments: the first time interaction video and the second time interaction video. These two videos are when participants discussed the first and second video they saw with the virtual agents. To test whether facial expressions at the beginning of the human-agent interaction is more indicative of the sense of rapport, we conducted the stepwise liner interaction again but entered AU 4, AU 9, AU 10 and AU 12 from the first and second interaction separately. The resulting model shows that display of AU 9 when participants first interacted with the agent is a significant predictor of...
the rapport established through the 2 interactions (F=11.92, p=.001, Beta=-.31).

7. Discussion

In this paper, we investigated the role of facial expression plays in rapport. Results from analysis of human-human interactions and human-virtual-human interactions show that facial expressions of disgust (or the lack of) indicated by the display of AU 9 and AU 10 are significant predictors of lack of rapport. The less disgust one displays, the more rapport one feels. Contrary to previous findings that people often use presence of smile as an indicator of rapport [5], we did not find smile (activation of AU 12 only, AU 6 was not included in the analysis) to be a significant predictor of rapport. There are literatures suggest that smile is not necessarily a reliable indicator of the quality of interaction. During social interaction, smile could mean genuine amusement and happiness as well as embarrassment, frustration and nervousness [25] [26].

Instead of using human expert to hand code facial expressions in the videos, we used a Computer Expression Recognition Toolbox (CERT) to automatically code the action units. We found the output of CERT to be rather accurate. For example, CERT output showed that when participants discussed the videos of positive emotional valance, they smiled more (more activity of AU 12), compared to when they discuss the video of negative emotional valance. This is consistent with previous findings that people do not frown more when talking about angry events but smile more talking about happy events [27]. One limitation of the current version of CERT is that it doesn’t output estimation of AU 6, which is widely considered to be part of the “Duchenne smile”.

Among human dyads, rapport can be conceptualized as a phenomenon occurring on three levels: the emotional, the behavioral, and the cognitive. Emotionally, we feel a harmony, a flow. Cognitively, we share an understanding with our conversation partner. Behaviorally, there is a convergence of movements with our conversational partner. Future work can further the current investigation to different levels of rapport. For example, facial expression may be more critical to the emotional rapport.

One limitation of the current work is that the Rapport Agent does not provide facial expression as part of the contingent feedback. Prior research show that people tend to mimic each other’s facial expression in a dyad interaction, even when they are not consciously aware of the other person’s facial expression [27][28]. Lack of facial expressions from the Rapport Agent may have forced participants to turn to other nonverbal channel to maintain positivity, such as head nods, as the interaction moved along. Paired sample t-test paring action units displayed in first interaction and second interaction (from the human-virtual-human study) did not show significant reduction of action unit activity. However, our results do show that facial expression displayed at the beginning of the interaction is more indicative of the sense of rapport than the ones displayed towards the end of the interaction. Further analysis of participants other nonverbal behavior such as gaze and head nods may help shed more light on this issue.

Another limitation of current work is that it does not address the contingency of facial expression in rapport. Work conceptualizing the nature of rapport has put great emphasis on the importance of interactional synchrony [10] [4]. Our early work on rapport inducing virtual agent also support this view [13]. Future work includes incorporating contingent feedback in response to human facial expression for the Virtual Rapport Agent.

Work presented in this paper show that people do display different facial expressions when engaged in high or low rapportful interactions. This is the first step towards informing the Virtual Rapport Agent about the human counterparts’ sense of rapport and adapting the agent’s behavior in response.

8. Acknowledgment

This work was sponsored by the U.S. Army Research, Development, and Engineering Command. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References


