Understanding and Predicting Bonding in Conversations Using Thin Slices of Facial Expressions and Body Language

Natasha Jaques[1], Daniel McDuff[2], Yoo Lim Kim[3], and Rosalind Picard[1]

[1] MIT Media Lab, Cambridge MA 02139, USA,
{jaquesn, picard}@media.mit.edu,
http://affect.media.mit.edu/
[2] Affectiva, Waltham MA 02453, USA
daniel.mcduff@affectiva.com
[3] Wellesley College, Wellesley MA 02481, USA
ykim9@wellesley.edu

Abstract. This paper investigates how an intelligent agent could be designed to both predict whether it is bonding with its user, and convey appropriate facial expression and body language responses to foster bonding. Video and Kinect recordings are collected from a series of naturalistic conversations, and a reliable measure of bonding is adapted and verified. A qualitative and quantitative analysis is conducted to determine the non-verbal cues that characterize both high and low bonding conversations. We then train a deep neural network classifier using one minute segments of facial expression and body language data, and show that it is able to accurately predict bonding in novel conversations.

Keywords: Facial expressions, body language, bonding, rapport

1 Introduction

The most effective conversationalists do not simply smile, nod, and mirror their partner; instead, they are adept at sensing non-verbal cues and adapting to the other person's state. If an intelligent virtual agent (IVA) could be designed with this level of emotional intelligence, it could dynamically adapt its interaction style to the needs of the user. Such an endearing and empathetic IVA would have a wide range of applications, from intelligent tutoring, to human-robot interaction, to helping individuals who struggle with social interaction.

In this study we show that using facial expression and body language data from one-minute segments of a conversation (a.k.a. *thin slices*), a machine learning classifier can be trained to predict whether a novel person will experience bonding up to twenty minutes later. While it has been shown that humans have the ability to predict similar outcomes from such thin slices of an interaction [1], training computer algorithms to predict bonding using this data is a novel contribution. Data is collected unobtrusively using cameras and Microsoft Kinects while participants engage in free-form conversations. Bonding is assessed empirically using a measure adapted from the Working Alliance Inventory [2]; we show that it is strongly related to conversation quality and rapport.

To provide insight into the data we have collected and the features extracted, we provide both a qualitative and quantitative analysis of facial expression and skeletal joint position features related to bonding. We also suggest ways that an IVA could learn to synthesize the appropriate non-verbal responses based on interaction context, and provide insight into the type of non-verbal behaviors that may arise in situations in which a person is either extremely frustrated with an interaction, or deeply engaged.

2 Related Work

A body of work has shown that using only thin slices (less than five minutes) of video of a person's non-verbal cues, human judges can predict everything from therapy outcomes to job performance [1]. Since computer algorithms have successfully predicted conversational outcomes like stress and engagement using audio data [3], it is possible that an IVA could use thin slices of facial expressions and body language to predict whether it is bonding with its user.

Non-verbal cues such as facial expressions and body language are a rich source of information about a person's mental state, and as such, there has been a great deal of research on how to detect, interpret, and display them. Although a thorough survey of all such work is impossible here, we refer the interested reader to a recent meta-analysis of the state of the art in automatic facial-expression recognition [4]. Automatic analysis of body-language has also been explored. For example, Avola and colleagues [5] developed a system that uses Kinect data to compute features of gestural strokes, and Yang and colleagues used motion capture data to show that friendly conversational dyads had a higher degree of correlation in body language gestures [6].

Most relevant to our work is research on bonding and rapport, which has been investigated in the context of the contingency (e.g. [7]) or mirroring (e.g. [8]) between the VA's behavior and the user. Detailed models of rapport have also been developed [9]. Other research has investigated which facial expressions generated by an agent led to the most rapport with its users [10].

3 User Study

Data were collected from a study in which participants conversed while being recorded with cameras, microphones, and Microsoft Kinects. To conceal the true nature of the study and ensure participants could act naturally, participants were told the purpose of the study was to train computer algorithms to read lips. They were instructed to stay within view of the recording devices, but not to over-emphasize their lip movements[4], and to keep the conversation flowing as naturally as possible. The interaction lasted for approximately 20 minutes, after which participants completed a post-study survey and were debriefed about the

[4] Even if some participants did speak with exaggerated lip movements, this would not affect our later analysis.

study's true purpose. All procedures were approved by the MIT IRB. In total we had 30 participants (13 male, 17 female) divided into 15 conversation dyads. There was variety across participants in terms of age (M = 40.0, SD = 15.3), occupation, ethnicity, and socioeconomic status.

The post-study survey contained a *Perception of Interaction* questionnaire similar to that of [11], in which participants gave Likert-scale ratings of their partner on a number of attributes, and completed the Bonding subscale of the Working Alliance Inventory (B-WAI). The WAI was developed to measure the degree of collaboration and trust between a therapist and their client; the bonding subscale is specifically designed to measure positive personal attachment, including "mutual trust, acceptance, and confidence" (p. 224) [2]. The scale was adapted for our study by removing three items irrelevant for short conversations between strangers (17, 21, and 36). The language of some other items was slightly modified to fit the interaction of a conversation; e.g. item 29 was changed to read "I had the feeling that if I said or did the wrong things, my partner would stop *talking* with me" rather than "working with me". Most items were unmodified. Typical items included "My partner and I understood each other", and "I felt uncomfortable with my partner".

4 Methods

Facial expression extraction. Automated software (Affdex - Affectiva, Inc.) [12] was applied to the videos to obtain confidence scores (from 0 to 100) indicating the presence of facial expressions. These included twelve facial action units from the Facial Action Coding System (FACS) [13], as well as smiles, lip corner pulls, seven expressions of emotion, and three axes of head pose (pitch, yaw and roll). After removing portions of the interaction in which the participant's face was not tracked, and downsampling each signal to 1 Hz to ensure smooth estimates, we obtained facial expression data for 13,714 seconds of conversation.

Skeletal joint extraction. Microsoft Kinects were used to gather data about the X (horizontal), Y (vertical) and Z (depth) position of participants' joints, including the head, neck, thumbs, finger tips, four positions on each limb, and three positions on the spine. To clean this data we removed portions of the interaction in which a second body was tracked, and 4s segments in which the derivative was more than two standard deviations above the mean in any axis (X, Y or Z) (which is often due to the Kinect briefly losing track of the participant). After removing noise, minutes of the interaction that were missing more than 60% of the data were discarded, due to the unreliability of the signal during this period. The joint data was then aligned with the video data. Finally, we applied a z-score normalization to the data from each axis of each joint, which reduces effects due to the Kinect being placed in slightly different locations for different participants.

Machine learning classification. To train our machine learning model, we extracted features from each minute of conversation for each participant and their partner. From the skeletal data, we computed five features for each joint's

X, Y, and Z positions: the mean, std. dev., max. of the abs. derivative, mean derivative, and max. of the abs. second derivative. These features provide information about the position, degree of movement, speed of movement, direction of movement, and sharpness of movement (acceleration), respectively. For facial expressions, we computed the sum, mean, and std. dev. of each feature, telling us the amount, degree, and variability in expression. In total we obtained 375 joint and 143 facial expression features for each of 532 minutes of conversation.

Each minute was assigned a binary classification label, based on whether it belonged to a conversation with high or low bonding (scores were split based on the median B-WAI). The data were then randomly partitioned into training, validation, and testing sets. Data from each participant were assigned to only one set. Thus, the testing set represents completely novel, held-out data.

To reduce the number of features, we used Correlation-based Feature Selection (CFS) [14]. CFS chooses a subset of features that are both strongly predictive of an outcome variable (in this case bonding), but also have low correlation with the rest of the features in the subset (are not redundant). CFS was applied only to the training data, to avoid contaminating the testing data. Neural network models were then trained on the CFS features using Google's TensorFlow library [15]. Both single-layer and deep architectures were explored, and parameters were tuned using the validation set.

5 Results

In this section we will provide evidence establishing the reliability of the modified B-WAI, and give examples of the type of data we have collected and ways in which it can be used to detect bonding. A quantitative analysis of the differences in facial expressions and body language between participants with high and low bonding will be provided. Finally, we show that machine learning can be applied to these features to accurately predict bonding up to 20 minutes later.

Reliability of the bonding scale. The following analysis relies on B-WAI as an aggregate measure of bonding, rapport, and participants' perceptions of their conversations as warm, comfortable, and enjoyable. To examine how well B-WAI captures these characteristics, we tested the correlations between it and eight self-reported Likert-scale ratings of conversation quality (see Table 1). We see that B-WAI is related to participants' ratings of their partner as *interesting*, *charming*, *friendly*, and *funny*, and inversely related to their ratings of *distant* and *annoying*. After applying a Bonferroni correction, the relationships between B-WAI and *interesting*, *annoying*, and *distant* remained significant, suggesting that B-WAI is strongly related to participants' perceived conversation quality.

Qualitative analysis. In this section we provide examples of the facial expression and body language data we have collected, showing that the interaction between the two participants is highly relevant to bonding. For example, Figure 1 plots five minutes of facial expressions which occurred between the participant who experienced the lowest bonding in our study, P_l (top), and her partner (bottom). Although P_l began the interaction with frequent smiling, in the portion

Table 1: Pearson's r correlations between B-WAI and conversation quality. Bolded measures are significant after performing a Bonferroni correction.

Measure	r	p
Interesting	**.6912**	**<.001**
Charming	.4342	.021
Friendly	.3806	.038
Funny	.3736	.046
Engaging	.1104	.561

(a) Positive correlations

Measure	r	p
Distant	**-.6207**	**<.001**
Annoying	**-.5549**	**.001**
Awkward	-.2589	.167

(b) Negative correlations

of the interaction plotted in Figure 1, she shows expressions of sadness as she is discussing a highly personal topic. Instead of responding empathetically, her partner continues to smile and smirk. Eventually P_l becomes angry, and afterwards simply stops emoting; for the rest of the conversation, she shows little or no facial expressions whatsoever. This interaction underlines the importance of designing an IVA to both detect emotional cues, and display the appropriate response at the right time. Further, it could suggest that an emotionally intelligent agent may need to treat a sudden suppression of affect as a potential warning sign of an upset or frustrated user.

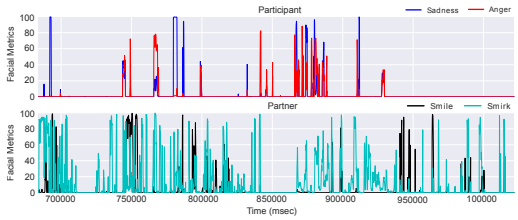


Fig. 1: Five minutes of facial expressions from the conversation with the least bonding, in which the participant's partner fails to respond empathetically.

While displaying the appropriate emotional cues in response to an unhappy user can be considered a minimum requirement of an emotionally intelligent VA, promoting a high degree of bonding and rapport can be much more subtle and complex. Figure 2 plots the Z position of the Spine Mid joint for the two participants in the conversation with the highest bonding. The distance maintained between the participants reveals a high degree of synchrony, suggesting they are highly attentive and responsive to each others' movements.

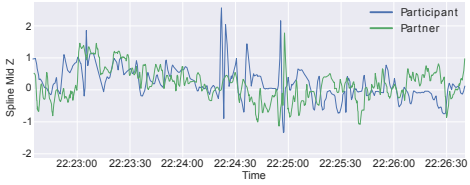
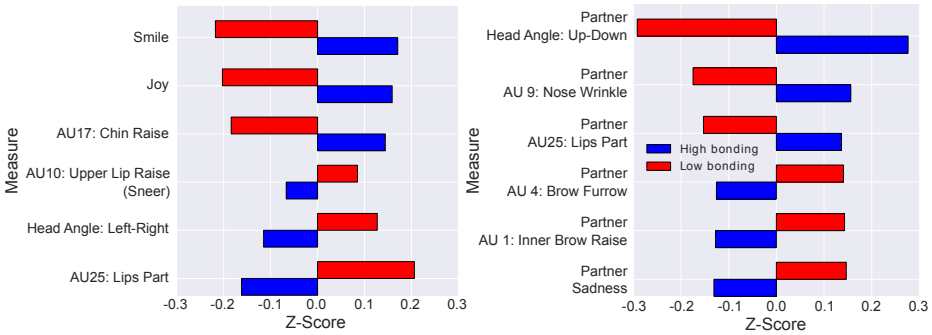


Fig. 2: Spine Mid Z for the participant with the highest bonding and her partner.

Quantitative analysis. In this section we will establish what kinds of facial expression and skeletal position features are relevant to bonding, and discuss design implications for an IVA. To begin, we analyzed which facial expression behaviors are more frequent in conversations with high vs. low bonding, by computing the difference in average *z-score* between both groups. Figure 3 shows the three features that had the greatest difference for both high and low bonding conversations, for the participant themselves, and their partner[5]. T-tests with a Bonferroni correction were used to assess whether high and low bonding conversations differed significantly on these features, and all of them reached significance at the $\alpha = .05$ level.



(a) Participant (b) Partner

Fig. 3: The participant and partner's facial expressions with the largest differences between conversations with high bonding (blue) and conversations with low bonding (red). If the Z-score is below zero, it means the behavior was less frequent in this group's conversations relative to the overall average.

Figure 3 reveals some expected trends. When the participant experiences bonding, she is more likely to smile, express joy, and raise her chin. When she feels that bonding is low, she is more likely to sneer and shake her head (see *Head angle left right* in Figure 3a). In terms of the partner's behavior, frequent nodding (*Head angle up down*) and nose wrinkling is associated with higher bonding. Nose wrinkling is often detected when someone is laughing, which has been found to be both deferential and endearing [16]. Conversely, negative displays of emotion by the partner appear to hinder bonding; frequent brow furrows, inner eyebrow raises, or expressions of sadness are associated with lower bonding. An intriguing thing to note, however, is that bonding is not symmetric. A participant who did not enjoy an interaction could score very low on the bonding scale, even though her partner felt fine about the conversation and scored relatively high (and indeed this does occur). Therefore asymmetric effects can occur, such as with the *lips part* AU (frequently detected when a person is speaking).

Although Figure 3 provides some interesting insights, without accounting for interaction context it can only give an incomplete picture of facial expressions

[5] The participant is the one who completes the B-WAI about their partner.

and bonding. Therefore we investigate how the contingency between conversational partners' expressions differs with bonding. We computed the Pearson's r correlations between each participant's facial expressions and their partner's, for conversations with high bonding (r_h) and low bonding (r_l). The difference between these coefficients was then computed as $r_{diff} = r_h - r_l$, and plotted in Figure 4. Blue locations in the grid correspond to behaviors that occurred together more frequently in high bonding conversations; red locations occurred more in low bonding conversations[6]. We also tested the correlation with the partner's behavior both 1s and 5s later. The results were similar, therefore we choose to focus on the behaviors that occur together in the same second, since neural processing of facial expressions occurs on the order of 100ms [17].

Figure 4 reveals several interesting patterns[7]. Bonding is likely to be lower if the partner is smiling or joyful while the participant is shaking her head. If the participant smiles while the partner is parting her lips the conversation is likely to have higher bonding, perhaps because the participant is enjoying what the partner is saying. Smiling at the right time appears to be important; bonding tends to be lower when the partner smiles or expresses joy in response to the participant's lip corner depressor or brow furrow. An interesting result is that there is little difference in the correlation between mutual smiling behavior in conversations with high and low bonding. This may suggest that mutual smiling is such a ubiquitous behavior that it can occur even when bonding is low.

Not only does Figure 4 provide insight into micro-interactions that can be used to detect bonding, it could also allow an IVA to synthesize appropriate facial responses. Consider the heatmap scores as probabilities that the agent could use when deciding what expression to display. If the user tilts her head (see the row *Head angle roll*), then the probability of the agent raising its outer eyebrows should be high, and the probability of it shaking its head should be low or almost zero. This approach is likely to be more effective than simple mirroring, because it captures the appropriateness of the expression in context.

A similar analysis is applied to the joint data collected with the Microsoft Kinect. After performing CFS feature selection as described in Section 4, we were left with a total of 69 non-redundant[8] joint features. For each of these, we computed the *information gain*, which can be interpreted as the reduction in uncertainty about one variable obtained after observing another [19]. Essentially, information gain tells us which features are most predictive of bonding. The five features with the highest information gain are listed in Table 2.

These joint features reveal that the partner's movements in the Z direction (towards or away from the participant) are highly related to whether the participant experiences bonding. The features relate to the position of the partner's

[6] Note again that bonding is not symmetric and neither is the matrix in Figure 4; it is computed based on the participant's perception of bonding, not her partner's.

[7] There are several strong differences in inner eyebrow raising, however this AU can be associated with either sadness or happiness, making it difficult to interpret [18].

[8] After CFS, two body part features that are highly correlated (for example, the left and right hips) will be represented by only one of the pair (e.g. the right hip).

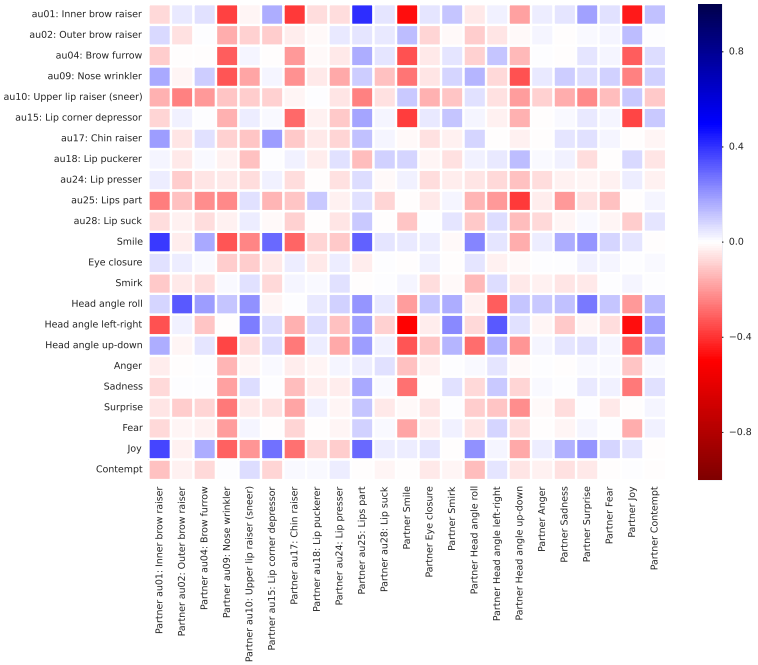


Fig. 4: The heatmap shows the difference in correlation coefficient ($r_{diff} = r_h - r_l$) between conversations with high bonding (r_h) and low bonding (r_l). Blue tiles represent a correlation that is more strongly positive in high bonding conversations, while red represents a correlation more prevalent in low bonding.

whole body, such as the spine, hips, and knees. Since these features describe the acceleration, variability, and speed of movement, a larger degree of movement of the partner's whole body may be more indicative of a high bonding conversation. Perhaps in conversations in which the partner is engaged and attentive, this enthusiasm is displayed by larger and more animated whole-body movements.

The synchrony between body language in conversation dyads must also be considered. As in the previous section, we computed the Pearson's r correlation between the participant's movements and her partner's in conversations with high and low bonding[9]. Interestingly, the speed and acceleration of whole body movements are highly correlated in conversations with high bonding. Correlations in acceleration in the Z direction are in some cases quite large; for example, for the knees, $r(308) = .5226, p < .001$, hips, $r(308) = .4578, p < .001$, and spine base, $r(308) = .4465, p < .001$. This suggests that in high-bonding conversations, the partner tends to closely mirror the sharpness of the participant's movements towards or away from her. This provides supporting evidence for the hypothesis generated in the previous section, that there is a great deal of synchrony in terms

[9] A similar heatmap was generated, but there is insufficient space to show it here.

Table 2: The skeletal joint features with the highest information gain. All features are significantly correlated with bonding after applying a Bonferroni correction.

Feature	Info. gain	Pearson's r	p
Partner SpineBaseZ sd	0.1695	0.4190	<.001
Partner HipRightZ sd	0.1541	0.4146	<.001
Partner KneeLeftZ max abs acc	0.1219	0.3505	<.001
Partner HipRightZ max abs deriv	0.1091	0.3712	<.001
Partner HipRightZ max abs acc	0.1091	0.3712	<.001

of whole body movements in pairs with high bonding. Agents that can mirror whole body movements (e.g. [8]) may be highly effective at facilitating bonding.

Predicting bonding in novel conversations. Using one-minute slices of the facial expression and body language features described above, we trained a series of neural network models to predict bonding, as explained in Section 4. We found that a deep architecture with 2 layers of 300 and 12 hidden nodes[10] led to the highest validation accuracy, of 64.7% (AUC=.678). Using this model, we obtained an accuracy of 85.87% and an AUC of .931 on the held-out test data, showing we can accurately predict bonding in novel conversations. Note that 66.3% of the samples in the test set belong to high-bonding interactions, so this result is almost 20% better than the baseline majority-class classifier (always guessing the most frequent class).

While these results are promising, they should be interpreted with caution given the small size of the testing dataset (N=92, representing data from 6 participants). While the validation accuracy still exceeded the majority-class baseline of 60.78% for the validation set, it was notably lower than the test accuracy. This is likely due to the random partitioning process and the small size of the datasets. Nevertheless, because the test set comprises novel users from which the classifier has never accessed data, this serves as a proof-of-concept that it is possible for an IVA to use data collected unobtrusively from a camera and Kinect to detect whether it is bonding with a new user during each minute of the conversation. Such fine-grained sensitivity to the user's perceptions could allow an IVA to dynamically adapt to improve bonding throughout the interaction, just like an excellent human conversationalist.

6 Conclusions and Future Work

We have shown that facial expression and body language features can allow an IVA to detect whether or not it is bonding with its user. We also presented a matrix, learned from human high and low bonding interactions, that could allow an IVA to generate the appropriate facial expressions and body language in response to user behavior. We have shown that a machine learning classifier can be trained to predict whether a person will experience high or low bonding, given only a one-minute slice of facial expression and body language data. This

[10] The other parameter settings were: learning rate = .01, batch size = 20, L2 regularization β = .01, no dropout.

information can be gathered unobtrusively with a camera and Kinect, making the classification system potentially highly useful to a future IVA.

As future work, the next step is to analyze the audio data, for prosody, emotional tone, and speaking turns. There are also many ways in which the modeling of the data could be improved. For example, a time-series analysis technique such as a Hidden Markov Model [19] could be employed to infer the participant's mental state (bonding or not) throughout the interaction, and the joint positions could be further abstracted into higher level gestures, as described by Avola et al. [5]. Even without these improvements, this work has contributed novel fundamental elements enabling the crafting of future agents with which human partners will bond.

References

1. N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.," *Psych. b.*, vol. 111, pp. 256, 1992.
2. A. Horvath and L. Greenberg, "Development and validation of the working alliance inventory.," *Journal of counseling psychology*, vol. 36, no. 2, pp. 223, 1989.
3. A. Pentland, "Social dynamics: Signals and behavior," in *Int. Conf. on Developmental Learning*, 2004, vol. 5.
4. M. Valstar et al., "Meta-analysis of the first facial expression recognition challenge," *Systems, Man, and Cybernetics*, vol. 42, no. 4, pp. 966–979, 2012.
5. D. Avola et al., "Human body language analysis: A preliminary study based on kinect skeleton tracking," in *ICIAP*, pp. 465–473. 2013.
6. Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *Multimedia*, vol. 16, no. 6, pp. 1766–1778, 2014.
7. J. Gratch et al., "Creating rapport with virtual agents," in *IVA*, 2007, pp. 125–138.
8. S. Kahl and S. Kopp, "Modeling a social brain for interactive agents: integrating mirroring and mentalizing," in *IVA*, 2015, pp. 77–86.
9. Zhao, Papangelis, and Cassell, "Towards a dyadic computational model of rapport management for human-virtual agent interaction," in *IVA*, 2014, pp. 514–527.
10. J. Wong and K. McGee, "Frown more, talk more: effects of facial expressions in establishing conversational rapport with virtual agents," in *IVA*, 2012, pp. 419–425.
11. R. Cuperman and W. Ickes, "Big five predictors of behavior and perceptions in initial dyadic interactions," *J. Pers. Soc. Psych.*, vol. 97, no. 4, pp. 667, 2009.
12. D. McDuff et al., "Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit," in *CHI*. ACM, 2016, pp. 3723–3726.
13. P. Ekman and W. Friesen, "Facial action coding system," 1977.
14. M. A. Hall, *Correlation-based Feature Subset Selection for Machine Learning*, Ph.D. thesis, University of Waikato, Hamilton, New Zealand, 1998.
15. M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.
16. Robert R Provine, *Laughter: A scientific investigation*, Penguin, 2001.
17. H. Meeren, C. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *PNAS*, vol. 102, 2005.
18. C. Kohler et al., "Differences in facial expressions of four universal emotions," *Psychiatry research*, vol. 128, no. 3, pp. 235–244, 2004.
19. K. P. Murphy, *Mach. learning: a probabilistic perspective*, MIT press, 2012.