Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures

Rana El Kaliouby and Peter Robinson
Computer Laboratory
University of Cambridge
Cambridge CB3 0FD, U.K.
{*rana.el-kaliouby, peter.robinson*}*@cl.cam.ac.uk*

Abstract

This paper presents a system for inferring complex mental states from video of facial expressions and head gestures in real-time. The system is based on a multi-level dynamic Bayesian network classifier which models complex mental states as a number of interacting facial and head displays, identified from component-based facial features. Experimental results for 6 mental states groups– agreement, concentrating, disagreement, interested, thinking and unsure are reported. Real-time performance, unobtrusiveness and lack of preprocessing make our system particularly suitable for user-independent human computer interaction.

1. Introduction

The human face provides an important, spontaneous channel for the communication of a wide array of mental states. Enabling man-machine interfaces to recognise and use the information conferred by this rich modality has gained significant research interest over the last few years. Facial expressions are used as conversation enhancers, to communicate feelings, show empathy and acknowledge the actions of other people [9]. Facial expressions also communicate cognitive mental states– often referred to as *complex* mental states– such as confused, thinking, and interested [2, 3]. These cognitive mental states occur more frequently in everyday interactions than their basic counterparts [18].

Despite the importance of complex mental states in interpreting and predicting the actions of others [20], facial expressions are almost always studied as a manifestation of basic emotions. The majority of existing automated facial expression analysis systems either attempt to identify basic units of muscular activity in the human face (action units or AUs) based on the Facial Action Coding System (FACS) [10], or only go as far as recognising the set of basic emotions [14, 6, 19, 8, 17, 7].

This paper describes a system for inferring complex mental states from video of facial expressions and head gestures in real-time. The challenge in automatically reading complex mental states from the face, stems from 3 key characteristics that make them essentially different from the simpler basic emotions. First, while basic emotions are arguably identifiable solely from facial action units, complex mental states additionally involve asynchronous information sources such as purposeful head gestures and eye-gaze direction [1]. Secondly, whereas basic emotions are identifiable from a small number of frames or even stills, complex mental states can only be reliably discerned by analysing the temporal dependencies across consecutive facial and head displays. Displays indicate different mental states when perceived with respect to preceding ones versus in isolation [11]. Thus modelling complex mental states involves multi-level temporal abstractions: at the highest level, mental states typically last between 6-8 seconds [3]. Head and facial displays can last up to 2 seconds, while at the lowest level, action units last tenths of seconds. Finally, whereas basic emotions have distinct facial expressions that are exploited by automated classifiers, finding facial and head displays relevant to complex mental states continues to be an active and challenging research problem [18, 1].

Based on those characteristics, we identify a number of requirements for the feature extraction and classifier methodology that we adopt. The classifier should: 1) be dynamic 2) deal with multiple interacting processes and 3) be able to model multi-level temporal abstractions. The feature extraction approach should be resilient to substantial rigid head motion, whilst being able to identify purposeful facial expressions and head gestures. In addition, because expert domain knowledge is not available, feature selection is needed to find optimal facial and head displays relevant to each mental state.

We describe two principle contributions: a system for inferring complex mental states from facial expressions and head gestures in real-time, as well as the optimal subset of facial and head displays that are most relevant in identifying the different mental states. Our system is built around a multi-level dynamic Bayesian network (DBN) classifier which models complex mental states as a number of interacting facial and head displays, identified from component-based facial features. Real-time performance, unobtrusiveness and lack of preprocessing make our system particularly suitable for spontaneous user-independent man-machine contexts.

The rest of the paper is organised as follows: in the next section we summarise related work, followed by an overview of our system (Section 3). Feature extraction, facial and head display recognition is discussed in Section 4, while Section 5 presents the dynamic Bayesian network models for complex mental states. Section 6 reports experimental results, before Section 7 concludes the paper.

2. Related work

We begin our review of related work with Garg et al's approach to multimodal speaker detection [4, 12] as this provides the inspiration for our present work. In their work, asynchronous audio and visual cues are fused along with contextual information and expert knowledge within a DBN framework. DBNs are a class of graphical probabilistic models which encode dependencies among sets of random variables evolving in time, with efficient algorithms for inference and learning. DBNs have also been used in unsupervised learning and clustering of facial displays [13]. Hidden Markov models (HMMs), the simplest kind of DBNs, are used by Lien et. al [14] to recognise facial AUs. Cohen et. al [5] use hierarchical HMMs to automatically segment an arbitrary long video sequence into different expression segments. Other classifier methodologies that have been applied to facial expression analysis include static ones such as Bayesian network classifiers that classify single frames into an emotion class (e.g. Cohen et. al [6]). Likewise, support vector machines have been used to classify feature point displacements compared to a neutral frame, into an emotion class [15].

While numerous approaches to feature extraction exist, those meeting the real-time constraints required for man-machine contexts are of particular interest. Methods such as principal component analysis and linear discriminant analysis of 2D face models (e.g. Padgett and Cottrell [16]), can potentially run in real-time but require initial pre-processing to put images in correspondence. Features based on facial point displacements are also common and and have shown validity when compared to manual FACS coding (e.g. Cohn et al. [7], the authors [15], and Pantic and Rothkrantz [17]). Tian et al. [19] use a combination of motion, shape and color descriptors to describe a number of face components such as the mouth and eyebrows.

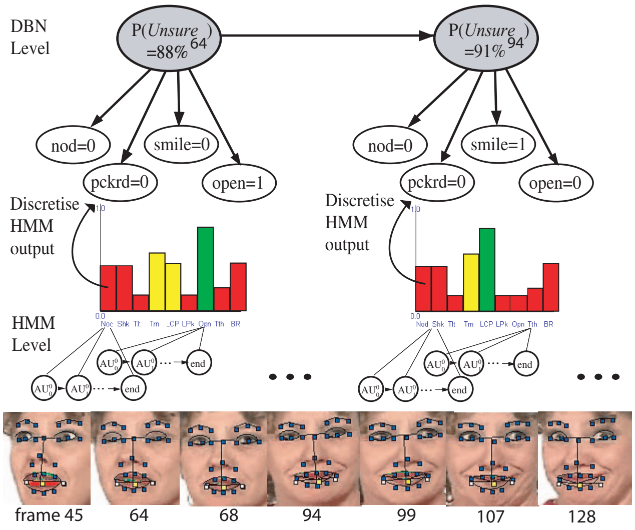


Figure 1: Overview of the multi-level system for inference of complex mental states from facial expressions and head gestures in real-time.

3. Overview

Our system (Fig. 1) involves three levels on progressively longer time scales: action unit analysis, facial and head display recognition, and mental state inference. A real-time facial feature tracker locates and tracks 24 facial features from video. The feature points define motion, color and shape descriptors for various face components (e.g. mouth). The descriptors are first normalised against head motion and then mapped to corresponding head or facial actions. HMM filters classify unseen sequences of actions into head and facial displays in real-time. The output likelihoods from the HMM classifiers are quantised and used as input to the DBN to infer the underlying complex mental state. The decision to model the HMM level separately rather than part of the DBN was taken to make the system more modular. For our purposes the two approaches have the same computational complexity.

Commodity hardware is used, such as a commercial digital camcorder placed near the user's monitor and connected to a standard PC. We assume a full frontal view of the face, but take into account variations in head pose inherent in video-based interaction. Videos are captured at 30 frames per second. While learning is done offline, inference is done in real-time by temporally abstracting each level of the system such that a classification per frame is not necessary. For example, based on empirical observations, head and facial actions occur over 200 millisecond intervals, while displays span between 0.6 to 1.2 seconds. Each level is implemented as a sliding window to make it possible to run the system for an indefinite duration.

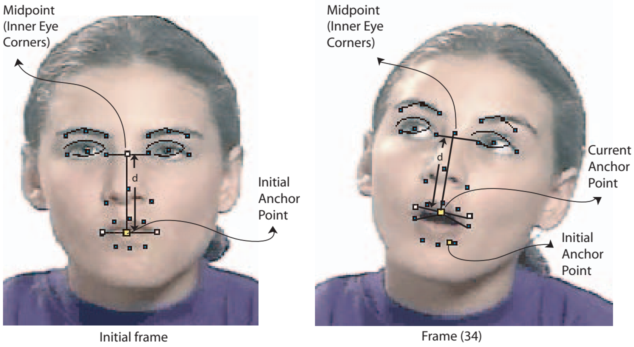
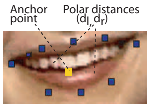
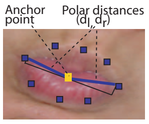
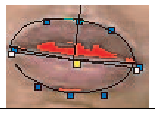
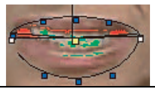
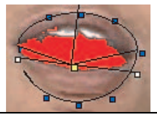


Figure 2: The "anchor" point in initial and subsequent frame of a video.

Table 1: List of mouth facial actions supported

Facial Action	Feature	Description
Lip Pull		$\frac{(d_{l_t}+d_{r_t})-(d_{l_0}+d_{r_0})}{(d_{l_0}+d_{r_0})} \geq k$
Lip Pucker		$\frac{(d_{l_t}+d_{r_t})-(d_{l_0}+d_{r_0})}{(d_{l_0}+d_{r_0})} \leq -k$
Lips Part		$aperture + teeth \simeq 0$
Mouth Stretch		$teeth \geq t$
Jaws drop		$aperture \geq a$

k, a and t are empirically determined

4. Facial and head action analysis

Twenty four facial landmarks are detected using a face template in the initial frame, and their positions tracked across the video. Head actions are identified from pose estimation points. Head pitch (up or down) is determined from the vertical displacement of the nose tip. Head yaw (turn) is given by the ratio of left to right eye widths, while a head roll (tilt) is given by the slope of the two inner eye corners. The sequences on which we tested include yaw, roll and pitch as large as 50, 30 and 50 degrees respectively.

To identify facial actions, we extract component-based facial features based on motion, geometry and color descriptors such as those in Tian et al. [19]. Component-based facial features are particularly suitable for a real-time video system, in which motion is inherent and places a strict upper bound on the computational complexity of methods used in order to meet time constraints.

We modify the components to account for out-of-plane head motion as follows: We imagine that the initial frame in the sequence is a reference frame attached to the head of the user. On that frame, let (X_p, Y_p) be an "anchor" point around which the head rotates. The point is the 2D projection corresponding to the imaginary point around which the head rotates in 3D space. As shown in Fig. 2, the anchor point is initially defined as the midpoint between the two mouth corners when the mouth is at rest, and is at a distance d from the line joining the two inner eye corners l. In subsequent frames the point is measured at distance d from l, after accounting for head turns. The anchor point is resilient to head rotations along the three axes, and is normalised against the distance between the two eye corners to account for scale variations.

The mouth is represented by a polygon connecting eight feature points. For every frame, the polar distance and angle are calculated with respect to the anchor point. Probability distribution functions based on luminance represent aperture and teeth pixels inside the mouth polygon. To remove the effects of variation in scale between image sequences in face size, all parameters are computed as ratios of the current values to that in the initial frame. In the case of a non-neutral initial frame, the polar angle is used to approximate the initial mouth state. The facial actions described by the mouth components are listed in Table 1. The lip corner pull and puckered are determined by the magnitude and direction of change of the polar distance and angle. The lips part, jaw drop and mouth stretch are discerned by the ratio of aperture (shown in red in Table 1) to teeth (shown in green) pixels. In addition, eyebrow components given by inner, center and outer feature points depict upper facial actions such as an eyebrow raise.

Facial and head actions are converted into symbol sequences and input into left-to-right HMM classifiers to identify facial expressions and head gestures. Each HMM is modelled as a temporal sequence of action units. For instance, a head nod display is a series of alternating head up (AU53), head down (AU54) movements, while a persistent, unidirectional head tilt display is a sequence of tilt actions (AU55 and/or AU56). A smile consists of an onset, peak, offset of lip corner pull (AU12, AU6+12). Each HMM is implemented as one of three topologies: 4-state, 3-symbol HMMs are used with head nods, head shakes and mouth displays, 2-state, 7-symbol HMMs represent tilt and turn displays, while a 2-state, 2-symbol HMM models an eyebrow raise.

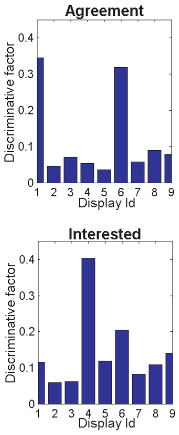


Figure 3: Discriminative ability of facial and head displays for 6 complex mental states. Display Ids are as follows, 1:nod, 2:jaw drop, 3:pucker, 4:raise, 5:shake, 6:lip pull, 7:mouth stretch, 8:tilt, 9:turn.

5. DBNs for complex mental states

We first analyse the discriminative power of head and facial displays for various complex mental states as there is little documentation in the literature on the facial "signatures" of such states. The discriminative power of display d for mental state m is determined by the difference in $P(d|m)$ and $P(\bar{d}|m)$. It follows that if d were a strong discriminator of m, the power function would approach 1. Fig. 3 summarises the results of analysing the discriminative power of 9 head and facial displays for 6 different complex mental states. The strongest discriminator was an eyebrow raise for *interested* (0.404) followed by a head nod in *agreement* (0.345). The analysis verifies that, on their own, facial expressions and head gestures are weak classifiers that do not capture underlying complex mental states. Bayesian networks, including DBNs have successfully been used as an ensemble of classifiers, where the combined classifier performs much better than any individual one in the set [12].

Each mental state is modelled as a separate DBN, allowing the system to be in more than one mental state at a time. This is particularly useful for mental states that are not mutually exclusive (e.g. *thinking* and *concentrating*). In our initial attempt at the DBN structure, all supported head and facial displays were included, making no assumptions about which displays contribute the most (or least) to particular mental states. We then implemented sequential backward selection that finds for each mental state the optimal subset of observation nodes $\{d_0, d_1, ...d_K\}$ from N supported displays, such that the discriminative ability F of the DBN model m, given by

$$F = \sum_{n=1}^{K} (\left|P(\bar{d_n}|m) - P(d_n|m)\right|), \qquad (1)$$

is maximised. Using these optimal subsets we build a DBN model specific to each mental state. The advantages of using only the most relevant features for the DBN model structure include: 1) reducing model dimensions without impeding performance of the learning algorithm, and 2) improving the generalisation power of each class by filtering irrelevant features.

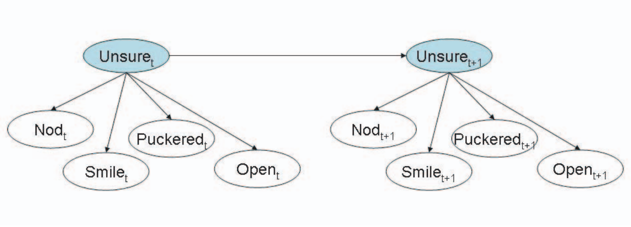


Figure 4: DBN model for *unsure* at consecutive instances.

Figure 4 illustrates the DBN model for *unsure*. Each node is a variable. The hidden (shaded) mental state node equals 1 whenever the user is unsure, and 0 otherwise. It influences four observation nodes (unshaded), which describe whether the user is nodding or not, smiling or not, puckering the lips or not, and has his/her mouth open or not. The arcs between the nodes are parameterised by conditional probability distributions that model dependencies between variables at time t. The arc between the two binary variables *unsure* and *nod*, for example, stores the two-by-two conditional probability table (CPT), $P(nod|unsure)$. We let B_ϕ denote the total set of CPT parameters. An additional arc has been placed between the hidden node unsure at consecutive times to encode temporal dependency between the variable in two slices of the network. The probability distribution is defined by a matrix of transition probabilities A and an initial state distribution π.

The set of network parameters θ=(B_ϕ, A, π) can be learned from a training data set using maximum likelihood training. Let e denote the hidden state and y denote the four observation nodes. Let $E_T = \{e_1, e_2, \ldots, e_T\}$ be the sequence of T hidden states and Y_T the corresponding sequence of observations. Then we have:

$$P(E_T, Y_T, \theta) = P(Y_T|E_T, B_\phi)P(E_T|A, \pi) \qquad (2)$$

When all the nodes are observed, the parameters B_ϕ can be determined by counting how often particular combinations of hidden state and observation values occur. The transition matrix A can be viewed as a second histogram which counts the number of transitions between the hidden state values over time. Inference is carried out using the classic forward-backward algorithm.

6. Experimental evaluation

We evaluate our system by considering classification performance for six complex mental state groups: *agreement*, *concentrating*, *disagreement, interested, thinking and unsure*. We use 106 videos from *Mind Reading* (MR), a computer-based guide to emotions [3]. Video durations vary between 5 to 8 seconds (SD=.45), recorded at 30fps. There are no restrictions on the head or body movement of actors in the video. To our knowledge MR is the only available, labelled resource with such a rich collection of mental states and emotions, albeit posed.

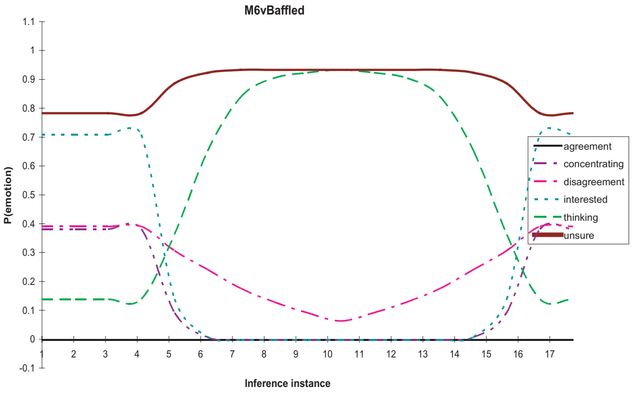


Figure 5: The likelihood of each mental state for a video of *baffled* plotted against time (inference instance).

Leave-5-out cross validation was used to split the 106 videos picked for evaluation into training and test sets. For each video, head and component-based facial actions are automatically extracted and input to the display HMMs. Six consecutive HMM likelihoods are quantised then input as evidence to the DBN classifiers. Each DBN classifier outputs the probability of the hidden mental state being true (i.e. 6 probability values per inference instance). Approximately 20 inferences are made in a video 6 seconds long, enabling the system to run in real-time. Since each level in the system is implemented using a sliding window, the 106 videos in effect, generate a total of 15738 and 2623 HMM and DBN samples respectively. Fig. 5 shows the likelihood of each mental state for a video of *baffled* plotted over time (inference instance). To determine if a video of T inference instances has been correctly classified, the most likely mental state m for the entire video is given by a minimisation error function E,

$$E = \min_{m} \sum_{t=1}^{T} (1 - P(m)), \tag{3}$$

Table 2: Breakdown of results for the mental state groups

Group	Mental State	#videos	%Correct
Agreement	Assertive	3	66.7
	Committed	5	100
	Convinced	4	100
	Decided	4	50
	Encouraging	3	100
	Sure	4	100
	Willing	2	100
	Total	**25**	**88.1**
Concentrating	Absorbed	4	100
	Concentrating	6	100
	Total	**10**	**100**
Disagreement	Contradictory	3	100
	Disapproving	5	40
	Discouraging	5	100
	Total	**13**	**80.0**
Interested	Asking	5	80
	Interested	5	100
	Total	**10**	**90.0**
Thinking	Brooding	3	66.7
	Calculating	4	75.0
	Choosing	5	100.0
	Fantasising	4	100
	Thinking	2	100
	Total	**18**	**88.9**
Unsure	Baffled	6	100
	Confused	6	83.3
	Puzzled	6	83.3
	Undecided	6	83.3
	Unsure	6	100
	Total	**30**	**90**
Overall Recognition Rate		**106**	**89.5**

An overall average classification rate of 89.5% was obtained. Table 6 summarises the breakdown of results for each of the mental state groups. In addition, a false positive rate F_m for mental state m (Table 6) is given by,

$$F_m = \frac{\text{Total number of videos falsely classified as } m}{\text{Total number of videos not } m} \tag{4}$$

A closer look at the results reveals a number of interesting points. First, onset frames occasionally portray a different mental state than the rest of the video. For example, the onset of *disapproving* videos were classified as *unsure*. Although this incorrectly biased the overall classification to *unsure*, one could argue that this result is not entirely incorrect and that the videos do indeed start off as *unsure*. Second, subclasses that do not clearly exhibit the class signature were easily misclassified. For example, *assertive* and *decided* videos in *agreement* were

Table 3: False positive rates for each mental state group

Group	#non-class	#false	%
Agreement	81	4	4.9
Concentrating	96	2	2.08
Disagreement	103	1	0.97
Interested	96	1	1.04
Thinking	88	0	0
Unsure	76	4	5.2

misclassified as *concentrating*, as they do not exhibit nods or smiles. Finally, we found that some mental states were "closer" to each other and could co-occur. For example, a majority of the *unsure* files scored high for *thinking* too. Further research is needed 1) to test the generalisation power of the system by evaluating a larger sample, which requires substantial investment in building a corpus of videos, and 2) to explore the relationship between complex mental states.

7. Conclusion

The two principle contributions of this paper are: 1) a multi-level DBN classifier for inferring complex mental states from videos of facial expressions and head gestures in real-time, and 2) insight into the optimal subset of facial and head displays most relevant in identifying different mental states. Those were used to drive the DBN structure. We reported promising results for 6 complex mental states. This paper serves as an important step towards integrating real-time facial affect inference in mainstream computing applications.

Acknowledgements

The authors would like to thank Professor Simon Baron-Cohen and his group at the Autism Research Centre, University of Cambridge for making Mind Reading available to our research.

References

[1] S. Baron-Cohen. How to build a baby that can read minds: Cognitive mechanisms in mindreading. *Current Psychology of Cognition*, 13(5):513–552, 1994.

[2] S. Baron-Cohen, A. Riviere, M. Fukushima, D. French, J. Hadwin, P. Cross, C. Bryant, and M. Sotillo. Reading the mind in the face: A cross-cultural and developmental study. *Visual Cognition*, 3:39–59, 1996.

[3] S. Baron-Cohen and T. H. E. Tead. Mind reading: The interactive guide to emotion. DVD Software (Jessica Kingsley Publishers), 2003.

[4] T. Choudhury, J. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In *International Conference on Pattern Recognition*, volume 3, pages 789–794, 2002.

[5] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding (CVIU) special issue on Face recognition.*, 91(1), 2003.

[6] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang. Learning bayesian network classifiers for facial expression recognition with both labeled and unlabeled data. In *IEEE conference on Computer Vision and Pattern Recognition*, volume 1, pages 595–604, 2003.

[7] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, 36:35–43, 1999.

[8] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[9] P. Ekman. *Human Ethology*, chapter About Brows: Emotional and conversational signals, pages 169–200. London: Cambridge University Press, 1979.

[10] P. Ekman and W. Friesen. *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978.

[11] R. el Kaliouby, P. Robinson, and S. Keates. Temporal context and the recognition of emotion from facial expression. In *Proceedings of HCI International Conference*, 2003.

[12] A. Garg, V. Pavlovic, and T. S. Huang. Bayesian networks as ensemble of classifiers. In *International Conference on Pattern Reconition*, volume 2, pages 20779–220784, 2002.

[13] J. Hoey and J. J. Little. Decision theoretic modeling of human facial displays. In *Proceedings of European Conference on Computer Vision*, 2004.

[14] J. Lien, A. Zlochower, J. Cohn, and T. Kanade. Automated facial expression recognition. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 390–395, 1998.

[15] P. Michel and R. el Kaliouby. Real time facial expression recognition in video using support vector machines. In *The International Conference on Multimodal Interfaces*, 2003.

[16] C. Padgett and G. Cottrell. Identifying emotion in static images. In *Processing of the second Joint Symposium of Neural Computatio*, volume 5, pages 91–101, 1995.

[17] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18:881–905, 2000.

[18] P. Rozin and A. B. Cohen. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of americans. *Emotion*, 3(1):68–75, 2003.

[19] Y.-L. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[20] H. Wellman. *The childs theory of mind*. Cambridge, MA: Bradford Books/MIT Press, 1990.