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

Rana El Kaliouby and Peter Robinson

Computer Laboratory
University of Cambridge
rana.el-kaliouby@cl.cam.ac.uk
peter.robinson@cl.cam.ac.uk

In this chapter, we describe a system for inferring complex mental states from a video stream of facial expressions and head gestures in real-time. The system abstracts video input into three levels, each representing head and facial events at different granularities of spatial and temporal abstraction. We use Dynamic Bayesian Networks to model the unfolding of head and facial displays, and corresponding mental states over time. We evaluate the system's recognition accuracy and real-time performance for 6 classes of complex mental states—*agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. Real-time performance, unobtrusiveness and lack of preprocessing make our system suitable for user-independent human-computer interaction.

1 Introduction

People exhibit and communicate a wide range of affective and cognitive mental states. This process of mind-reading, or attributing a mental state to a person from the observed behaviour of that person is fundamental to social interaction. Mind-reading allows people to make sense of other's actions within an intentional framework [1]. The majority of people read the minds of others all the time, and those who lack the ability to do so, such as people diagnosed along the autism spectrum, are at a disadvantage [2]. Beyond social interaction, there is growing evidence to show that emotions regulate and bias processes such as perception, decision-making and empathic understanding, in a way that contributes positively to intelligent functioning [8, 13, 23].

The human face provides an important, spontaneous channel for the communication of mental states. Facial expressions function as conversation enhancers, communicate feelings and cognitive mental states, show empathy and acknowledge the actions of other people [6, 15]. Over the past decade there has been significant progress on automated facial expression analysis (see Pantic and Rothkrantz [35] for a survey). The application of automated

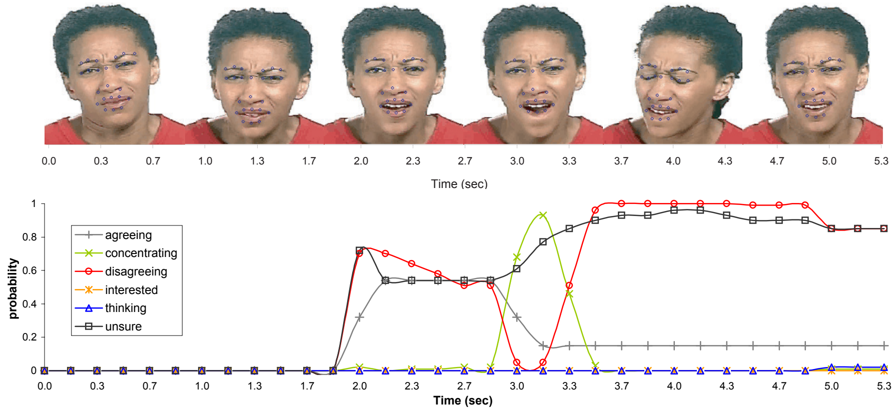


Fig. 1. Mental state inference in a video labelled as *discouraging* from the Mind Reading DVD [5]: (top) selected frames sampled every 1 s; (bottom) results of mental state inference. The overall probability of *disagreeing* is 0.75, a correct classification

facial expression analysis to human-computer interaction (HCI) however, is limited to basic, inconsequential scenarios. This is because the majority of existing 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) [16], or only go as far as recognizing the set of basic emotions [11, 12, 14, 29, 36, 39]. The basic emotions comprise only a small subset of the mental states that people can experience, and are arguably not the most frequently occurring in day-to-day interactions [38].

In this chapter, we describe a system for inferring complex mental states from a video stream of facial expressions and head gestures in real-time. The term complex mental states collectively refers to those mental states—both affective and cognitive—that are not part of the classic basic emotions, and which, as a result have not been addressed by the computer science research community. The system makes two principal contributions. First, it classifies different shades of complex mental state classes, and second, it does so from a video stream of facial events in real-time. Figure 1 shows the output of the system for a video labelled as *discouraging* from the Mind Reading DVD [5]. It is our belief that by building systems that recognize a wide range of mental states, we widen the scope of HCI scenarios in which this technology can be integrated.

2 Related Work

We begin our review of related work with Garg et al.'s approach to multimodal speaker detection [19] 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 Dynamic Bayesian Network (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 activity recognition and facial event analysis. Park and Aggarwal [37] present a DBN framework for analyzing human actions and interactions in video. Hoey and Little [22] use DBNs in the unsupervised learning and clustering of facial displays. Zhang and Ji [42] apply DBNs to recognize facial expressions of basic emotions. Gu and Ji [20] use DBNs to classify facial events for monitoring driver vigilance. Other classifiers that have been applied to facial expression analysis include static ones such as Bayesian Networks and Support Vector Machines that classify single frames into an emotion class [11, 32].

The input to the classifiers are features extracted from still or video sequences. 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., [34]), can potentially run in real-time but require initial pre-processing to put images in correspondence. Gabor wavelets as in Littlewort et al. [30] are feature independent but are less robust to rigid head motion and require extensive (sometimes manual) alignment of frames in a video sequence. The approach that we adopt for feature extraction is based on the movement of points belonging to facial features [12, 36, 32]. Facial analysis based on feature-point tracking compares favourably to manual FACS coding [12].

3 The Mind Reading DVD

Existing corpora of nonverbal expressions, such as the Cohn-Kanade facial expression database [26], are of limited use to our research since they only cover enactments of the classic basic emotions. Instead, we use the Mind Reading DVD [5], a computer-based guide to emotions, developed by a team of psychologists led by Professor Simon Baron-Cohen at the Autism Research Centre, University of Cambridge. The DVD was designed to help individuals diagnosed along the autism spectrum recognize facial expressions of emotions.

The DVD is based on a taxonomy of emotion by Baron-Cohen et al. [4] that covers a wide range of affective and cognitive mental states. The taxonomy lists 412 mental state concepts, each assigned to one (and only one) of 24 mental state classes. The 24 classes were chosen such that the semantic distinctiveness of the emotion concepts within one class is preserved. The number of concepts within a mental state class that one is able to identify reflect one's empathizing ability [3].

Out of the 24 classes, we focus on the automated recognition of 6 classes that are particularly relevant in a human-computer interaction context, and that are not in the basic emotion set. The 6 classes are: *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure.* The classes include affective states such as *interested*, and cognitive ones such as *thinking*, and encompass 29 mental state concepts, or fine shades, of the 6 mental states. For instance, *brooding*, *calculating*, and *fantasizing* are different shades of the *thinking* class; likewise, *baffled*, *confused* and *puzzled* are concepts within the *unsure* class.

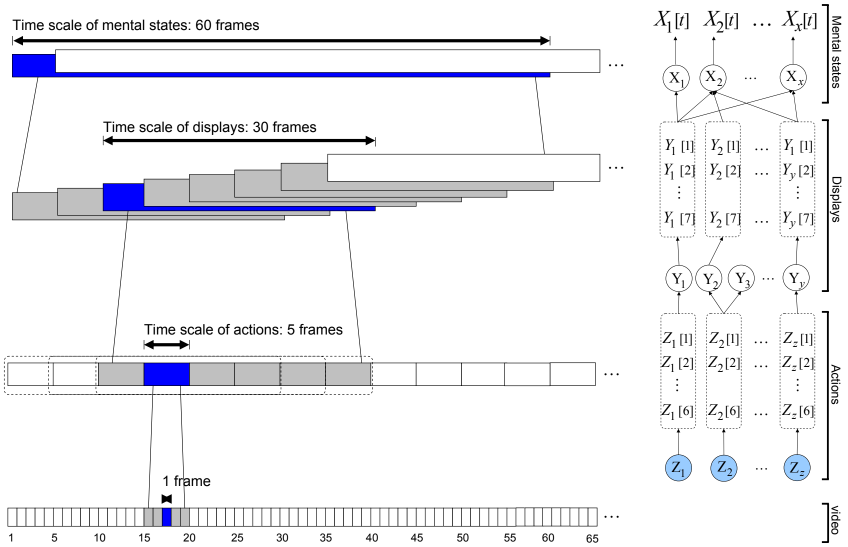
Each of the 29 mental states is captured through six video clips. The resulting 174 videos were recorded at 30 frames per second, and last between 5 to 8 seconds, compared to a mean duration of .67 seconds per sequence in the Cohn-Kanade database [26]. The resolution is 320×240. The videos were acted by 30 actors of varying age ranges and ethnic origins. All the videos were frontal with a uniform white background. The process of labelling the videos involved a panel of 10 judges who were asked 'could this be *the emotion name*?' When 8 out of 10 judges agreed, a statistically significant majority, the video was included. To the best of our knowledge, the Mind Reading DVD is the only available, labelled resource with such a rich collection of mental states, even if they are posed.

4 The Automated Mind-Reading System

A person's mental state is not directly available to an observer (the machine in this case) and as a result has to be inferred from observable behaviour such as facial signals. The process of reading a person's mental state in the face is inherently uncertain. Different people with the same mental state may exhibit very different facial expressions, with varying intensities and durations. In addition, the recognition of head and facial displays is a noisy process.

To account for this uncertainty, we pursued a multi-level representation of the video input, combined in a Bayesian inference framework. Our system abstracts raw video input into three levels, each conveying face-based events at different granularities of spatial and temporal abstraction. Each level captures a different degree of temporal detail depicted by the physical property of the events at that level. As shown in Fig. 2, the observation (input) at any one level is a temporal sequence of the output of lower layers.

Our approach has a number of advantages. First, higher-level classifiers are less sensitive to variations in the environment because their observations are the outputs of the middle classifiers. Second, with each of the layers being trained independently, the system is easier to interpret and improve at different levels. Third, the Bayesian framework provides a principled approach to combine multiple sources of information. Finally, by combining dynamic modelling with multi-level temporal abstraction, the model fully accounts for the dynamics inherent in facial behaviour. In terms of implementation, the system is user-independent, unobtrusive, and accounts for rigid head motion while recognizing meaningful head gestures.


Time scale of mental states: 60 frames
Time scale of displays: 30 frames
Time scale of actions: 5 frames
1 frame
1
5
10
15
20
25
30
35
40
45
50
55
60
65
…
$X_1[t]$ $X_2[t]$ … $X_x[t]$
X_1 X_2 … X_x
Mental states
Y_1[1] Y_2[1] … Y_1[1]
Y_1[2] Y_2[2] … Y_y[2]
Y_1[7] Y_2[7] … Y_y[7]
Y_1 Y_2 Y_3 … Y_y
Displays
Z_1[1] Z_2[1] … Z_z[1]
Z_1[2] Z_2[2] … Z_z[2]
Z_1[6] Z_2[6] … Z_z[6]
Z_1 Z_2 … Z_z
Actions
video

(a) Time scales at each level of the system. On level L a single event is shown in black. The input to this event is a sequence of events from level $L-1$ (shown in gray). A single action spans 5 frames (166 ms), a display spans 30 frames (1 s), and a mental state spans 60 frames (2 s).

(b) Matrix representation of the output at each level of the system.

Fig. 2. Multi-level temporal abstraction in the system

4.1 Extraction of Head and Facial Actions

The first level of the system models the basic spatial and motion characteristics of the face including the head pose. These are described by z facial actions $\mathbf{Z} = \{Z_1, \ldots, Z_z\}$ based on the FACS. Each action describes the underlying motion of an abstraction across multiple frames. Figure 3 summarizes the spatial abstractions currently supported by the model: head rotation along each of the three rotation axes (pitch, yaw and roll) and facial components (lips, mouth and eyebrows). For example, $Z_1[t]$ may represent the head pose along the pitch axis at time t; the possible values of Z_1 are $\{\text{AU53}, \text{AU54}, \textit{null}\}$ or the head-up AU, head-down, or neither respectively. To determine the time scale of head and facial actions, we timed the duration of 80 head-up and 97 head-down motions in head nod gestures, sampled from 20 videos by 15 people representing a range of complex mental states such as *convinced*, *encouraging* and *willing*. The movements lasted at least 170 ms, a result similar to that in the kinematics of gestures [9]. The system produces facial or head actions every 5 frames at 30 fps, or approximately every 166 ms.

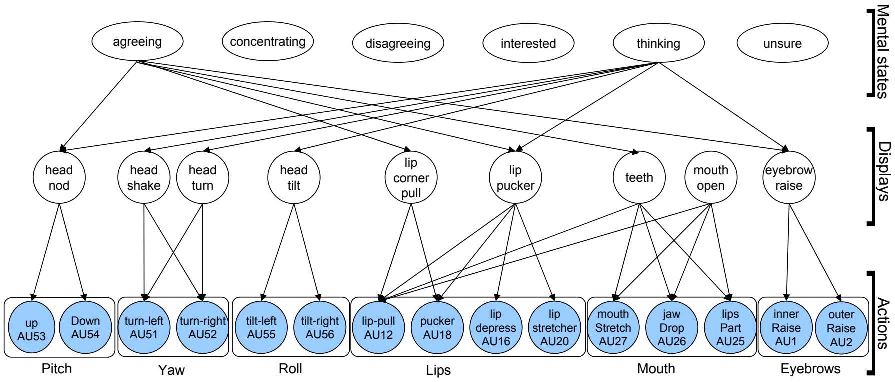
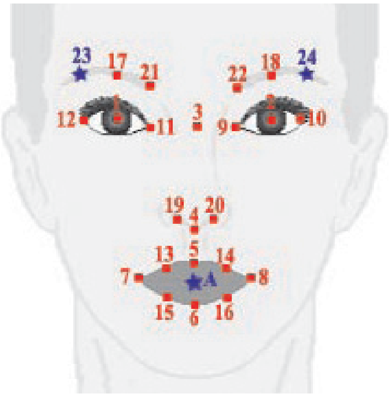


Fig. 3. A video stream is abstracted spatially into head pitch, yaw and roll actions, and lips, mouth and eyebrow actions. The actions are in turn abstracted into displays and mental states. The displays present in a model of a mental state are determined by a feature selection mechanism. For clarity, the displays for only two mental states are shown

For head and facial action extraction, feature points are first located on the face and tracked across consecutive frames using `FaceTracker` [18], part of Nevenvision's facial feature tracking SDK. Figure 4 describes the 2D model of the face used by the system, and how the head and facial AUs are measured. The motion of expression-invariant feature points over successive frames such as the nose tip, nose root, and inner and outer eye corners are used to extract head rotation parameters. This approach has been successfully used in a number of existing systems [33, 28, 39, 27]. A more accurate, but computationally intensive approach involves tracking the entire head region using a 3D head model [10, 17, 41]. Since our objective was to identify head actions automatically and in real-time, rather than come up with a precise 3D estimate of the head pose, a feature-point based approach was deemed more suitable than a model-based one. Facial actions are identified from motion, shape and colour descriptors derived from the feature points. The shape descriptors capture the deformation of the lips and eyebrows, while the colour-based analysis is used to extract the mouth actions (aperture and teeth).

4.2 Recognition of Head and Facial Displays

Head and facial actions are in turn abstracted into $y = 9$ head and facial displays $\mathbf{Y} = \{Y_1, \ldots, Y_y\}$. Displays are communicative facial events such as a head nod, smile or eyebrow flash. Each display is described by an event that is associated with a particular spatial abstraction as in the action level. Like actions, display events can occur simultaneously. $P(Y_j[t])$ describes the probability that display event j has occurred at time t. For example, Y_1 may



1. Head yaw	$\frac{\overline{P_9 P_{10}}}{\overline{P_{11} P_{12}}}$
2. Head pitch	$P_4[t] - P_4[t-1]$
3. Head roll	$\angle P_9 P_{11}$
4. Eyebrow raise	$\frac{(\overline{P_{11}P_{21}}+\overline{P_1P_{17}}+\overline{P_{12}P_{23}})_t}{(\overline{P_{11}P_{21}}+\overline{P_1P_{17}}+\overline{P_{12}P_{23}})_0}$
5. Lip pull	$\frac{(\overline{AP_7}+\overline{AP_8})_t-(\overline{AP_7}+\overline{AP_8})_0}{(\overline{AP_7}+\overline{AP_8})_0} > k$
6. Lip pucker	$\frac{(\overline{AP_7}+\overline{AP_8})_t-(\overline{AP_7}+\overline{AP_8})_0}{(\overline{AP_7}+\overline{AP_8})_0} < k$
7. Lips part	$\sum Aperture + \sum Teeth \approx 0$
8. Jaw drop	$\sum Aperture \geq \sum Teeth \geq a$
9. Teeth	$\sum Teeth \geq t$

Fig. 4. Extraction of head and facial actions: (left) the 25 fiducial landmarks tracker per frame; (right) action descriptors. P_i represents point i in the face model

represent the head nod event; $P(Y_1[t]|Z_1[1:t])$ is the probability that a head nod has occurred at time t given a sequence of head pitch actions. We timed the temporal intervals of 50 head-nod (AU53) and 50 head-shake gestures; a single display lasted 1 second on average. Accordingly, the time scale of a single display is 30 frames at 30 fps, or 6 actions. The output progresses one action at a time, i.e., every 166 ms.

To exploit the dynamics of displays, we use Hidden Markov Models (HMMs) for the classification of temporal sequences of actions into a corresponding head or facial display. Although defining the topology of an HMM is essentially a trial-and-error process, the number of states in each HMM were picked such that it is proportional to the complexity of the patterns that each HMM will need to distinguish; the number of symbols were determined by the number of identifiable actions per HMM. Accordingly, the head nod and head shake were implemented as a 2-state, 3-symbol ergodic HMM; episodic head turn and tilt displays as 2-state, 7-symbol HMMs to encode intensity, lip displays such as a smile, or pucker and mouth displays as in a jaw drop or mouth stretch, are represented by a 2-state 3-symbol HMM; the eye-brow raise as a 2-state, 2-symbol HMM. We decided to model the HMM level separately rather than part of the DBN to make the system more modular. For our purposes the two approaches have the same computational complexity.

4.3 Mental State Inference

Finally, at the topmost level, the system represents $x = 6$ mental state events $\{X_1, \ldots, X_x\}$. For example, X_1 may represent the mental state *agreeing*; $P(X_1[t])$ is the probability that *agreeing* was detected at time t. The probability $P(X_i[t])$ of a mental state event is conditioned on the most recently observed displays $\mathbf{Y}[1:t]$, and previous inferences of that mental state $P(X_i[1:t-1])$. We found that two seconds is the minimum time required for a human to reliably infer a mental state; video segments of less than 2 seconds result in inaccurate recognition results [25]. As shown earlier in Fig. 2, we chose to sample these 2 seconds using a sliding window of 30 frames, sliding it 6 times, 5 frames at a time. In terms of displays, the sliding window spans 1 display and progresses 6 times one display at a time.

Representation

We use DBNs to model the unfolding of head and facial displays, and corresponding mental states over time. DBNs are an appealing framework for complex vision-based inference problems. DBNs function as an ensemble of classifiers, where the combined classifier performs better than any individual one in the set [19]. They also incorporate multiple asynchronous cues within a coherent framework, and can model data at multiple temporal scales making them well suited to modelling hierarchically structured human behaviour.

To represent the x mental state classes, we decided to model each mental state as a separate DBN, where the hidden mental state of each DBN represents a mental state event. The event has two possible outcomes: it is true whenever the user is experiencing that mental state, and false otherwise. Having a DBN per class means that the hidden state of more than one DBN can be true; mental states that are not mutually exclusive or may co-occur can be represented by the system.

Like all probabilistic graphical models, a DBN is depicted by its structure and a set of parameters. The structure of the model consists of the specification of a set of conditional independence relations for the probability model, or a set of (missing) edges in the graph. The parameter set θ_i for mental state i is described in terms of an observation function, a state-transition function, and a prior. The observation function B_ϕ is parameterized by conditional probability distributions that model the dependencies between the two nodes. The transition function A encodes temporal dependency between the variable in two slices of the network. The prior π the initial state distributions. The model is given by its joint probability distribution:

$$P(X_i, \mathbf{Y}, \theta) = P(\mathbf{Y}|X_i, B_\phi)P(X_i|A, \pi)$$

4.4 Parameter Learning

When the data is fully observed and the network structure is known, Maximum Likelihood Estimation (MLE) can be used to estimate the parameters of a DBN. 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

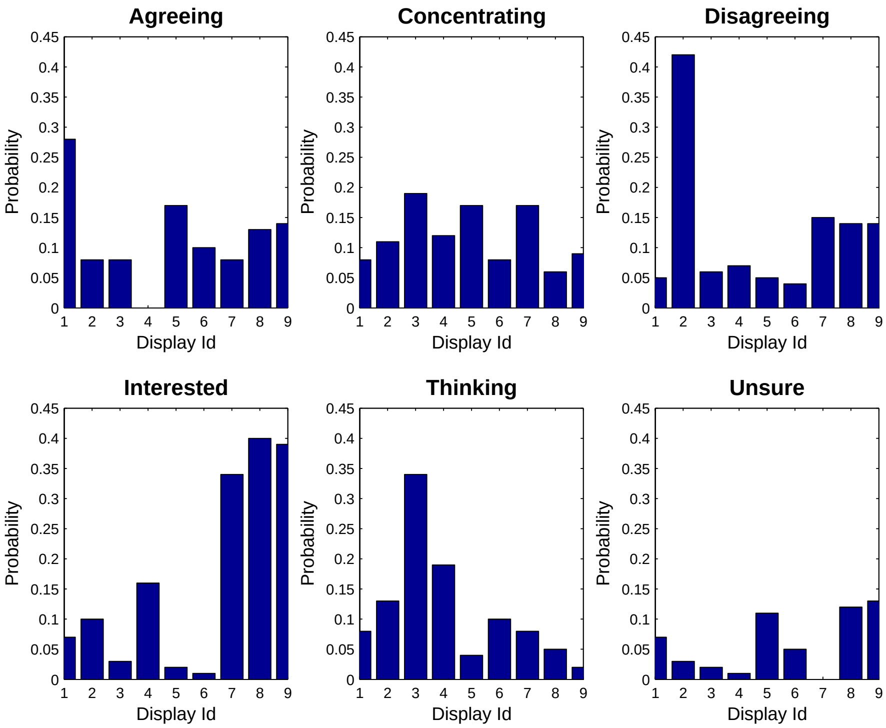


Fig. 5. Discriminative power of head and facial displays in complex mental states. Display Ids: 1:nod, 2:shake, 3:tilt, 4:turn, 5:lip-pull, 6:pucker, 7:mouth open, 8:teeth present, 9:eyebrow raise

histogram which counts the number of transitions between the hidden state values over time.

In addition to the above parameters, we define a heuristic H that quantifies the discriminative power of a display for a mental state: $H = P(Y_j|X_i) - P(Y_j|\overline{X}_i)$. The magnitude of H is an indication of which displays contribute the most (or least) to recognizing specific mental states. The sign depicts whether it increases or decreases the probability of the mental state. Figure 5 summarizes the discriminative power of head and facial displays for 6 different complex mental states.

A post-hoc analysis of the results of parameter estimation yields an insight into the facial expressions of complex mental states. The exercise is an important one given the little literature there is on the facial expressions of these states. The strongest discriminator was the head shake for *disagreeing* (0.42), followed by an eyebrow raise for *interested* (0.40). The analysis shows that single displays are weak classifiers that do not capture complex mental states, verifying the suitability of DBNs.

Table 1. Summary of model selection results. Column i summarizes how the probability of mental state i is affected by observing evidence on each of the displays. Row j depicts the effect of observing display j on the probability of each of the mental states

	agreeing	concentrating	disagreeing	interested	thinking	unsure
head nod	+0.28	-0.08	-0.05	-0.07	-0.08	-0.07
head shake		-0.11	+0.42		-0.13	+0.04
head tilt		-.019	-0.06		+0.34	
head turn					+0.18	
lip corner pull	+0.17	-0.17				-0.1
lip pucker	-0.10				+0.1	+0.06
mouth open	-0.13	+0.07	-0.14	+0.40		-0.05
teeth present	-0.14		-0.14	+0.39		-0.17
eyebrow raise	-0.08	-0.17	-0.15	+0.34	-0.08	

Model Selection

The results of parameter estimation show that the head and facial displays that are most relevant in discriminating mental states are not by necessity the same across mental states. This observation provided the motivation to implement model selection in search for the optimal subset of head and facial displays most relevant in identifying each of the mental states. Using only the most relevant features for the DBN structure reduces the model dimensions without impeding the performance of the learning algorithm, and improves the generalization power of each class by filtering irrelevant features.

Assuming the inter-slice topology is fixed, the problem of feature selection is an optimization one defined as follows: given the set of y displays $\mathbf{Y}$, select a subset that leads to the smallest classification error for videos in a test set of size S. Each video in the set yields T instances of mental state inference. The classification error per video per instance is $1 - P(X_i[t])$. Accordingly, the classification error of mental state i is given by the sum of the error over the T instances for all S videos:

$$e_i = \frac{1}{ST} \sum_{s=1}^{S} \sum_{t=1}^{T} (1 - P(X_i[t])) \tag{1}$$

We implemented sequential backward elimination [31] to find the optimal subset of observation nodes for each mental state. Features are removed recursively such that the classification error, e_i, of the DBN model is minimized. Note that the algorithm does not guarantee a global optima since that depends on the training and test sets used.

The results of sequential backward elimination are summarized in Table 1. A non-blank entry at cell (j, i) implies that display j is present in the DBN model of mental state i; the number is the value of the discriminative-power heuristic H of display j for mental state i. A positive value means that observing display j increases $P(X_i)$; a negative one means that observing display j

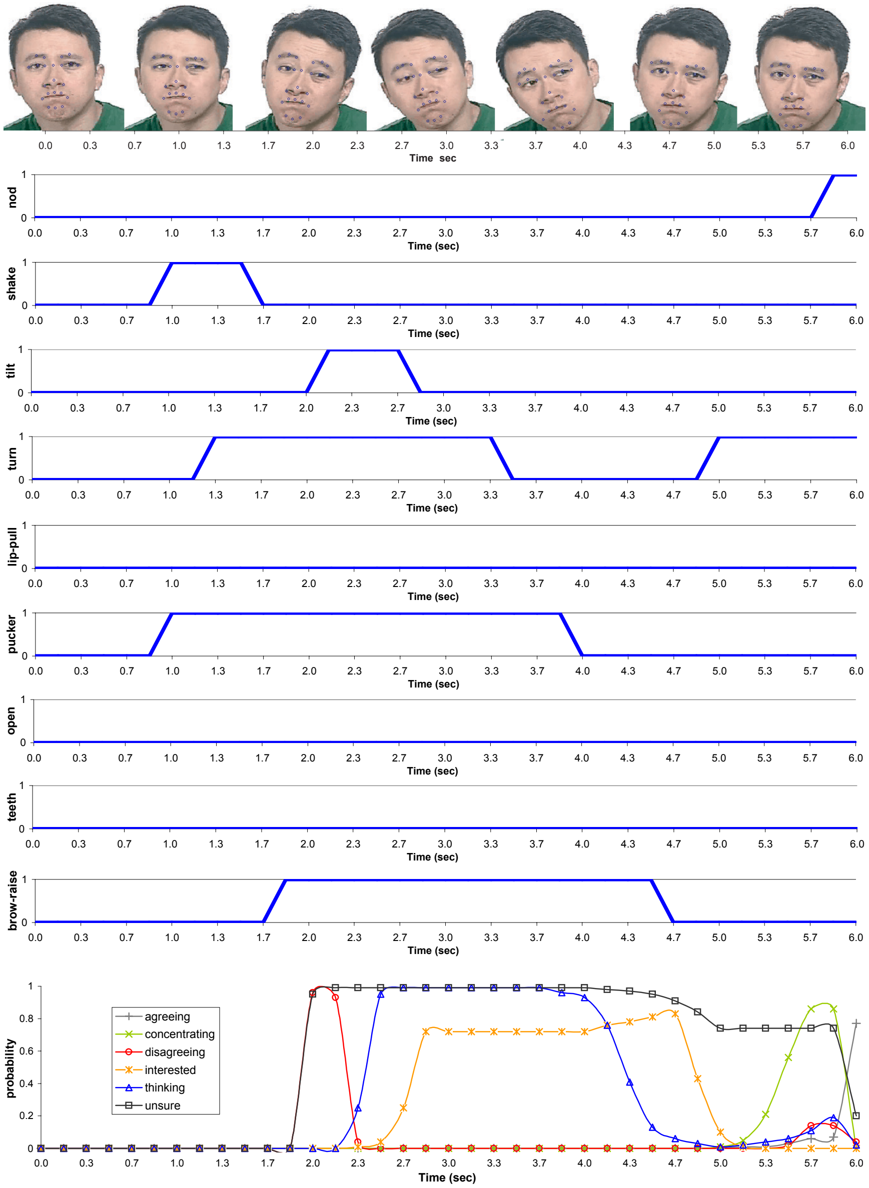
decreases that probability. The magnitude depicts the extent with which the probability will change. The columns summarize how each mental state is affected by observing evidence on each of the displays. For instance, the table predicts that an open mouth, teeth or eyebrow raise would increase the probability of *interested*, but a head nod would decrease it (assuming it was non-zero). The row depict the effect of observing a display on the probability of each of the mental states. For instance, observing a head shake would increase the probability of *disagreeing* and *unsure* but would decrease that of *concentrating* and *thinking*. Note that the table only provides a prediction; the actual behaviour of the DBNs will depend on the combination of displays recognized, their dynamics, and the probability of the previous mental states.

5 Recognition Accuracy

The accuracy is a measure of the classification performance of the system on a pre-defined set of classes. Those classes are *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. The objective of this experiment was to test how well the system performs when the 29 mental state concepts in each of the 6 classes are included. Each concept is represented by 6 videos from the Mind Reading DVD for a total of 174 videos. The challenge that the test posed is that while the concepts share the semantic meaning of the class they belong to, they differ in intensity, in the underlying head and facial displays, and in the dynamics of these displays. To the best of our knowledge, this is the first time different shades of a mental state are included in the evaluation of an automated facial expression analysis system.

5.1 Classification Rule

A classification rule is needed to determine whether or not the result of classifying each video in the test set is a correct one. The classification rule that we have used is a combination of the least-error rule with a threshold rule. The threshold rule was necessary because the least-error rule alone ignores the system's explicit representation of co-occurring mental states. The classification result for a video that is truth-labelled as i is a correct one if $e_i = e_{\min}$ or $e_i <= 0.4$, that is, if the class with the least-error matches the label of the video, or if on the whole the inferences result in the label of the video at least 60% of the time. Figure 6 shows an example display recognition and mental state inference in a 6-second long video labelled as *undecided* from the Mind Reading DVD. Throughout the video, a number of asynchronous displays that vary in duration are recognized: a head shake, a head tilt, a head turn, a lip pucker, and an eye-brow raise. The displays affect the inferred mental states over time as shown in the figure. The error value e is shown for each of the classes over the entire video as in (1). Since *undecided* belongs to the *unsure* class, and *unsure* scored the least error (and also meets the threshold), this is an example of a correct classification.

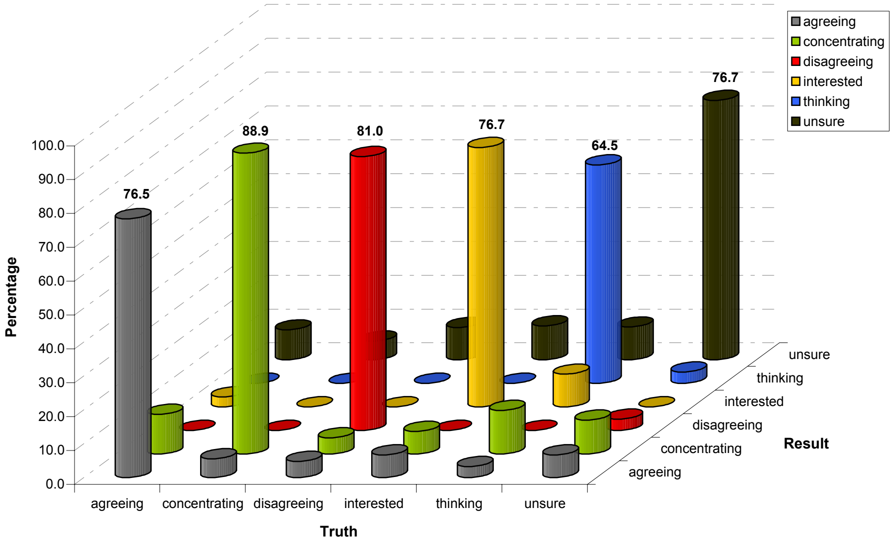


	agreeing	*concentrating*	*disagreeing*	*interested*	*thinking*	*unsure*
error value e	0.96	0.90	0.91	0.61	0.52	0.11

Fig. 6. Trace of display recognition and mental state inference in a video labelled as *undecided* from the DVD [5]: (top) selected frames from the video sampled every 1 second; (middle) head and facial displays; (bottom) mental state inferences for each of the six mental state classes and corresponding table of errors. Since the least error is *unsure* and *undecided* belongs to the *unsure* class, this is a correct classification

5.2 Results

Out of the 174 videos, 10 were discarded because FaceTracker failed to locate the non-frontal face on the initial frames of the videos. We tested the system on the remaining 164 videos, which spanned 25645 frames or approximately 855 seconds. Using a leave-one-out methodology, 164 runs were carried out, where for each run the system was trained on all but one video, and then tested with that video. Note that chance responding is at 16.7% since this is effectively a 6-way forced choice procedure. Chance responding describes a classifier that picks a class at random, i.e., does not encode any useful information.



mental state	agreeing	concentrating	disagreeing	interested	thinking	unsure	TP %
agreeing	**26**	4	0	1	0	3	76.5
concentrating	1	**16**	0	0	0	1	88.9
disagreeing	1	1	**17**	0	0	2	81.0
interested	2	2	0	**23**	0	3	76.7
thinking	1	4	0	3	**20**	3	64.5
unsure	2	3	1	0	1	**23**	76.7
FP %	5.4	9.6	0.7	3.0	0.8	9.0	**77.4**

Fig. 7. Recognition accuracy: (top) 3D bar chart of results (bottom) confusion matrix. The last column of the matrix is the true positive (TP) or classification rate for each class; the last row yields the false positive (FP) rate. For a false positive rate of 4.7%, the overall recognition accuracy of the system is 77.4%

The results are summarized in the confusion matrix and 3D bar chart in Fig. 7. Row i of the matrix describes the classification results for mental state

i. Column i states the number of times mental state i was recognized. The last column states the true positive (TP) or classification rate for class i. It is given by the ratio of videos correctly classified as mental state i to the total number of videos truth-labelled as i. The bottom row yields the false positive (FP) rate for class i, computed as the ratio of videos falsely classified as i to the total number of videos truth-labelled as anything but i. In the 3D bar chart, the horizontal axis describes the classification results of each mental state class. The percentage that a certain mental state was recognized is given along the $z-$axis. The classification rate is highest for *concentrating* (88.9%) and lowest for *thinking* (64.5%). The false positive rate is highest for *concentrating* (9.6%) and lowest for *disagreeing* (0.7%). For a mean false positive rate of 5.1%, the overall accuracy of the system is 77.4%.

5.3 Discussion

The overall accuracy of the system (77.4%) and the classification rates of each of the 6 classes are all substantially higher than chance responding (16.7%). Unfortunately, it is not possible to compare the results to those of other systems since there are no prior results on the automated recognition of complex mental states. Instead we compare the results to those reported in the literature of automated recognition of basic emotions, and to human recognition of complex mental states.

The accuracy of automated classifiers of basic emotions typically range between 85–95% [35]. Although this is higher than the results reported here, it is somewhat expected since the basic emotions are by definition easier to identify than complex ones, especially in stimuli that is stripped out of context. From an engineering point of view, the automated recognition of complex mental states is a challenging endeavour compared to basic emotions. This is because basic emotions have distinct facial expressions that are exploited by automated classifiers, while the facial expressions of complex mental states remains an open research problem. In addition, the DVD was not developed with automation in mind, so the videos are technically challenging compared to existing facial expression databases in a number of ways:

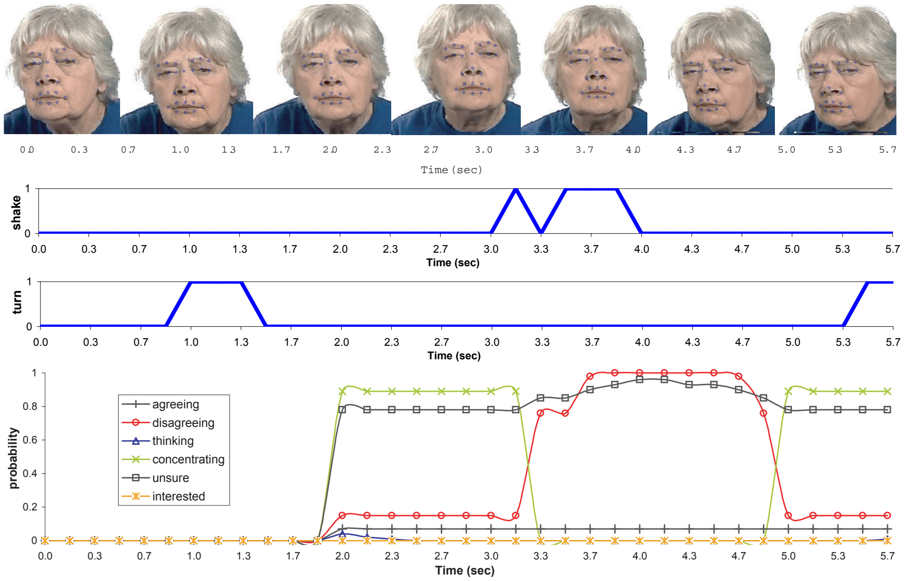
- Within-class variation
- Uncontrolled rigid head motion
- Multiple, asynchronous displays
- noisy evidence

Videos within a class vary along several dimensions including the specific mental states they communicate, the underlying configuration and dynamics of head and facial displays, and the physiognomies of the actors. In contrast, the stimuli used in training and evaluating existing automated facial analysis systems are typically more homogeneous, confined to a single prototypic expression of an emotion class. Hence, a video that varies substantially compared to other videos in the class along any of these dimension may end up being misclassified. For instance, only 60% of the videos labelled as *assertive* were correctly classified as *agreeing*. The rest were misclassified as *concentrating* or *unsure* since the underlying displays did not contain a head nod or a lip-corner pull (a smile) the most frequently observed displays in the *agreeing* class. The accuracy results then, will largely depend on the specific concepts that are picked for training and testing in each class and how different are their underlying displays. When the mental state concepts that share the underlying head/facial displays are only the ones picked for training and testing the system, the results reported are much higher. For example, in an earlier version of the system we reported an overall accuracy of 89.5% for 106 videos that cover 24 mental state concepts [24].

In terms of the underlying head and facial displays, there were no restrictions on the head or body movements of the actors, and there were no instructions given on how to act a mental state. Hence, the resulting head gestures and facial expressions are natural, even if the mental state is posed. In addition, while each video is given a single mental state label, it comprises of several asynchronous head and facial displays. Processing displays in context of each other by considering the transitions between displays, boosts the recognition results of humans for complex mental states [25]. Existing automated facial analysis systems of basic emotions, on the other hand, rely solely on facial expressions for classification and do not support the recognition of head gestures. Accordingly, the stimuli used in evaluating these systems is often restricted in terms of rigid head motion: the actors of these images or videos are either asked not to move their head, or are asked to exhibit very controlled head motion, and typically consists of a small number of frames limited to a single facial expression.

Finally, the head and facial display HMM classifiers are imperfect: displays may be misclassified or may pass undetected by the system altogether. Both cases will result in incorrect evidence being presented to the mental state DBNs. Depending on the persistence of the erroneous evidence, its location within the video, and its discriminative power, the resulting mental state inferences may be incorrect. Figure 8 shows an example of misclassification due to noisy evidence. The 5.7 second long video is labelled as *vigilant*, and is in the *concentrating* class. The output starts with a high probability of *concentrating*, which drops to 0 when a head shake is observed at 3.0 seconds. The head shake however, is a falsely detected display that persists for 1 second. At 5.0 seconds the head shake is no longer observed, and the probability of *concentrating* shoots up again. Unfortunately though, the effect of the head shake was such that *concentrating* did not score the least error and did not meet the 0.4 threshold and the video ended up being misclassified.

In a preliminary study [25] we show that human recognition of complex mental states from the Mind Reading DVD [5] is lower than that of the classic basic emotions, and reaches an upper bound of 71% for videos from the DVD. At 77.4%, the results of the automated mind-reading system are comparable to that of humans.



	agreeing	*concentrating*	*disagreeing*	*interested*	*thinking*	*unsure*
error value e	0.93	0.50	0.51	1.0	0.99	0.17

Fig. 8. Incorrect classification due to noisy evidence: (top) selected frames—sampled every 1 second—from a video labelled as *vigilant* from the DVD [5]; (middle) head and facial displays; (bottom) mental state inferences for each of the six mental state classes and corresponding table of errors. Note the effect of the false head shake on decreasing the probability of *concentrating*. The rest of the displays are not shown since there was nothing detected by the HMMs

6 Real-Time Performance

Real-time performance pertains to a system's ability to respond to an event without a noticeable delay. Executing in real-time is crucial since the idea is that applications adapt their responses depending on the inferred mental state of the user; it is pointless for an application to respond to a confused user long after she is no longer experiencing this mental state.

6.1 Objectives

The objective of this analysis is to quantify the real-time performance of the automated mind-reading system. The throughput and the latency are typically used to quantify the real-time performance of a vision-based system [40]. The **throughput** is the number of events that are processed per unit time. For the automated mind-reading system, the throughput translates to the number of mental state inferences made per second. The **latency** is defined as the time elapsed, or delay, between the onset of an event and when the system recognizes it. For the automated mind-reading system, the latency translates to the time it takes the system to infer the mental state, from the time a frame is captured.

6.2 Results

The processing time at each level of the system was measured on a Pentium IV (3.4 GHz) processor with 2 GB of memory. The results are summarized in Table 2. For feature point tracking, `Facetracker` runs at an average of 3.0 ms per frame of video at a resolution of 320×240 captured at 30 fps. The time taken to extract a single action was sampled over 180 function calls. On average, head-action function calls took 0.02 ms per frame depending on the amount of head motion in the frame; facial-action function calls lasted 0.01 ms per frame. In total, this level of the system executes at 0.09 ms per frame. The time taken to compute the probability of a head/facial display was also sampled over 180 invocations of the HMM inference algorithm. On average, a call to the HMM inference lasts 0.016 ms. Since there are nine displays implemented so far, this level of the system executes at 0.14 ms every five frames. Finally, the implementation of fixed lag smoothing of the six previous inferences using unrolled junction tree inference for a DBN with an average of seven nodes (one hidden mental state and six observation nodes) takes 6.85 ms per slice. Hence, this level executes at 41.10 ms for the six complex mental states.

Table 2. The processing time at each level of the automated mind-reading system (measured on a Pentium IV (3.4 GHz) processor with 2 GB of memory)

level	tracking	action-level	display-level	mental state-level	**total**
time (ms)	3.00	0.09	0.14	41.10	**44.33**

6.3 Discussion

To be deemed as real-time, the throughput of the system has to be at least six instances of mental states inferences per second to keep up with the input. This is because the DBNs are invoked every 5 frames at a capture rate of 30 frames per second. Also, the latency of the system has to be comparable to the latency of high-level facial expression recognition in humans, which ranges between 140–160 ms [7]. In our current implementation, the DBNs are the bottleneck of the system. Nonetheless, since 41.1 ms is less than 166 ms, the system runs in real-time. The total processing time for a frame is 44.34 ms. In terms of scalability, feature-point tracking, the extraction of head and facial actions and displays all run in linear time. At the mental state level, inference runs in polynomial time in the number of nodes [21].

7 Conclusion

The two principal contributions of this chapter are: 1) an automated system for inferring complex mental states, 2) a system that classifies the video input in real-time. The results also yield an insight into the optimal subset of facial and head displays most relevant in identifying different mental states. We reported promising results for the recognition accuracy and speed performance of 6 classes of complex mental states. Further research is needed to test the generalization power of the system by evaluating the system on a completely different previously unseen corpus of videos. The system we have presented serves as an important step towards integrating real-time facial affect inference in man-machine interfaces.

Acknowledgments

The authors would like to thank Professor Simon Baron-Cohen and Ofer Golan at the Autism Research Centre, University of Cambridge for making the Mind Reading DVD available to our research. This research was funded by the Computer Laboratory's Wiseman Fund, the Overseas Research Student Award, the Cambridge Overseas Trust, and Newnham College Studentship Research Award.

References

1. Simon 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. Simon Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, 1995.
3. Simon Baron-Cohen. *The Essential Difference: The Truth about the Male and Female Brain*. Perseus Publishing, 2003.
4. Simon Baron-Cohen, Ofer Golan, Sally Wheelwright, , and Jacqueline Hill. A New Taxonomy of Human Emotions. 2004.
5. Simon Baron-Cohen, Ofer Golan, Sally Wheelwright, and Jacqueline J. Hill. *Mind Reading: The Interactive Guide to Emotions*. London: Jessica Kingsley Publishers, 2004.
6. Simon Baron-Cohen, Angel Riviere, Masato Fukushima, Davina French, Julie Hadwin, Pippa Cross, Catherine Bryant, and Maria Sotillo. Reading the Mind in the Face: A Cross-cultural and Developmental Study. *Visual Cognition*, 3:39–59, 1996.
7. Magali Batty and Margot J. Taylor. Early Processing of the Six Basic Facial Emotional Expressions. *Cognitive Brain Research*, 17:613–620, 2003.
8. Antoine Bechara, Hanna Damasio, and Antonio R. Damasio. Emotion, Decision making and the Orbitofrontal Cortex. *Cereb Cortex*, 10(3):295–307, 2000.
9. Ray Birdwhistell. *Kinesics and Context*. University of Pennsylvania press, 1970.

10. Marco La Cascia, Stan Sclaroff, and Vassilis Athitsos. Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(4):322–336, 2000.
11. Ira Cohen, Nicu Sebe, Fabio G. Cozman, Marcelo C. Cirelo, and Thomas S. Huang. Learning Bayesian Network Classifiers for Facial Expression Recognition with both Labeled and Unlabeled Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 595–604, 2003.
12. Jeffrey F. Cohn, Adena J. Zlochower, James J. Lien, and Tokeo Kanade. Automated Face Analysis by Feature Point Tracking has High Concurrent Validity with Manual FACS coding. *Psychophysiology*, 36:35–43, 1999.
13. Antonio R. Damasio. *Descartes Error: Emotion, Reason and the Human Brain.* Putnam Sons: NY, 1994.
14. Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrance J. Sejnowski. Classifying Facial Actions. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 21(10):974–989, 1999.
15. Paul Ekman. *Human Ethology*, chapter About Brows: Emotional and Conversational Signals, pages 169–200. London: Cambridge University Press, 1979.
16. Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, 1978.
17. Murat Erdem and Stan Sclaroff. Automatic Detection of Relevant Head Gestures in American Sign Language Communication. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 1, pages 10460–10463, 2002.
18. FaceTracker. *Facial Feature Tracking SDK.* Neven Vision, 2002.
19. Ashutosh Garg, Vladimir Pavlovic, and Thomas S. Huang. Bayesian Networks as Ensemble of Classifiers. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 2, pages 20779–220784, 2002.
20. Haisong Gu and Qiang Ji. Facial Event Classification with Task Oriented Dynamic Bayesian Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 870–875, 2004.
21. Haipeng Guo and William H. Hsu. A Survey of Algorithms for Real-Time Bayesian Network Inference. In *AAAI/KDD/UAI Joint Workshop on Real-Time Decision Support and Diagnosis Systems*, 2002.
22. Jesse Hoey and James J. Little. Bayesian Clustering of Optical Flow Fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1086–1093, 2003.
23. Alice M. Isen. *Handbook of Emotions*, chapter Positive Affect and Decision Making, pages 417–435. Guilford Press, New York, 2000.
24. Rana el Kaliouby and Peter Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *2004 IEEE Workshop on Real-Time Vision for Human-Computer Interaction at the 2004 IEEE CVPR Conference*, 2004.
25. Rana el Kaliouby, Peter Robinson, and Simeon Keates. Temporal Context and the Recognition of Emotion from Facial Expression. In *Proceedings of HCI International Conference*, 2003.
26. Tokeo Kanade, Jeffrey Cohn, and Ying-Li Tian. Comprehensive Database for Facial Expression Analysis. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

27. Ashish Kapoor and Rosalind W. Picard. A Real-Time Head Nod and Shake Detector. In *Proceedings of the Workshop on Perceptive User Interfaces*, 2001.
28. Shinjiro Kawato and Jun. Ohya. Real-time Detection of Nodding and Head-shaking by Directly Detecting and Tracking the "Between-Eyes". In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 40–45, 2000.
29. James J. Lien, Adena Zlochower, Jeffrey F. Cohn, and Tokeo Kanade. Automated Facial Expression Recognition. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, 1998.
30. Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joel Chenu, Takayuki Kanda, Hiroshi Ishiguro, and Javier R. Movellan. Towards Social Robots: Automatic Evaluation of Human-robot Interaction by Face Detection and Expression Classification. In S. Thrun and B. Schoelkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, 2004.
31. Thomas Marill and David M.Green. On the Effectiveness of Receptors in Recognition Systems. *IEEE Transactions*, IT-9:11–27, 1963.
32. Philipp Michel and Rana el Kaliouby. Real Time Facial Expression Recognition in Video using Support Vector Machines. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 258–264, 2003.
33. Carlos Morimoto, Yaser Yacoob, and Larry Davis. Recognition of Head Gestures using Hidden Markov Models. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 461–465, 1996.
34. C. Padgett and G. Cottrell. Identifying Emotion in Static Images. In *Proceedings of the second Joint Symposium of Neural Computation*, volume 5, pages 91–101, 1995.
35. Maja Pantic and Leon J.M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22:1424–1445, 2000.
36. Maja Pantic and Leon J.M. Rothkrantz. Expert System for Automatic Analysis of Facial Expressions. *Image and Vision Computing*, 18:881–905, 2000.
37. Sangho Park and J.K. Aggarwal. Semantic-level Understanding of Human Actions and Interactions using Event Hierarchy. In *IEEE Workshop on Articulated and Non Rigid Motion at the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
38. Paul Rozin and Adam 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.
39. Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Recognizing Action Units for Facial Expression Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 23(2):97–115, February 2001.
40. Matthew Turk and Mathias Kolsch. *Emerging Topics in Computer Vision*, chapter Perceptual Interfaces. Prentice Hall PTR, 2004.
41. Jing Xiao, Toekeo Kanade, and Jeffrey F. Cohn. Robust Full Motion Recovery of Head by Dynamic Templates and Re-registration Techniques. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 163–169, 2002.
42. Yongmian Zhang and Qiang Ji. Facial Expression Understanding in Image Sequences Using Dynamic and Active Visual Information Fusion. pages 1297–1304, 2003.