Generalization of a Vision-Based Computational Model of Mind-Reading

Rana el Kaliouby and Peter Robinson

Computer Laboratory, University of Cambridge,
15 JJ Thomson Avenue, Cambridge UK CB3 0FD

Abstract. This paper describes a vision-based computational model of mind-reading that infers complex mental states from head and facial expressions in real-time. The generalization ability of the system is evaluated on videos that were posed by lay people in a relatively uncontrolled recording environment for six mental states—agreeing, concentrating, disagreeing, interested, thinking and unsure. The results show that the system's accuracy is comparable to that of humans on the same corpus.

1 Introduction

Existing human-computer interfaces are oblivious to the user's mental states and intentions, and as a result often respond inappropriately, e.g., by deciding to do irrelevant, computationally intensive tasks while a user is frantically working on a deadline. With the increasing complexity of human-computer interaction and the ubiquity of mobile and wearable devices, a new interaction paradigm is needed. In this paradigm, systems need to have socio-emotional intelligence to gather information autonomously about the user's state and to respond adaptively to it. "Theory of mind" or **mind-reading** [1]—the ability to attribute mental states to others by observing their behaviour—is a key component of socio-emotional intelligence in humans, and is equally important for natural user interfaces.

We have developed a computational model of mind-reading that infers mental states from head gestures and facial expressions in a video stream in real-time. The principal contribution of our system is the inference of complex mental states, states of mind that are not part of the set of basic emotions. These encompass affective states such as *interested*, and cognitive states that reveal mental processes such as *thinking* [3]. The automated inference of complex mental states has received almost no attention compared with the automated recognition of basic emotions. By supporting a wider range of mental states beyond the basic emotions, our system has widened the scope of human-computer interaction scenarios in which automated facial analysis systems can be integrated.

In this paper, we test the generalization ability of the computational model of mind-reading on videos that were posed by lay people in a relatively uncontrolled environment, having trained the system on carefully-composed videos from the Mind-Reading DVD [2][1]. We emphasize that training and testing are

[1] Video examples that demonstrate how our system generalizes can be found at `http://www.cl.cam.ac.uk/~re227/demo/`

J. Tao, T. Tan, and R.W. Picard (Eds.): ACII 2005, LNCS 3784, pp. 582–589, 2005.

carried out using different corpora, as opposed to using sampling methods on the same corpus (e.g., cross-validation) since sampling often introduces a bias in the results because of the similarity of recording conditions and actors on a single corpus. Evaluating the generalization ability of supervised systems is important: ultimately we need train the system on some (limited) data-set then deploy it in different scenarios with many users without having to re-train or calibrate it.

2 Related Work

Over the past decade, significant progress has been made with automated facial expression analysis (FEA) (a survey can be found in Pantic and Rothkrantz [15] and in Fasel and Leuttin [6]). The majority of systems are either concerned with the recognition of basic emotions (happy, sad, angry, disgusted, surprised and afraid) or with the automated coding of facial actions. There are two recent exceptions. Gu and Ji [7] present a facial event classifier for driver vigilance (inattention, yawning and state of falling asleep). Kapoor *et al.* [11] devise a multi-modal probabilistic framework for the recognition of interest and boredom. Only a few FEA systems have been evaluated with regards to how well they generalize across different corpora. This is mainly because the collection, filtering and labelling of videos is a labour intensive and time-consuming task. Littlewort *et al.* [12] test their system in recognizing the basic emotions when trained on the Cohn-Kanade database [10] and tested on the Pictures of Facial Affect [4] and vice versa. Even though both corpora contain similar stimuli—prototypic facial representations of the basic emotions and no rigid head motion—an average of 60% was reported compared with 95% when the system was trained and tested on the same corpus. In Michel and el Kaliouby [13], the system's accuracy dropped from 87.5% when trained and tested on the Cohn-Kanade database, to 60.7% when tested with users who were oblivious of the prototypic faces of basic emotions. A similar divergence of results is reported in Tian *et al.* [17] and Pardas *et al.* [16], emphasizing the importance of generalization in FEA systems.

3 Computational Model of Mind-Reading

A person's mental state is not directly available to an observer; instead it is inferred from nonverbal cues such as facial expressions. We present a novel approach to mental state representation based on the theory of mind-reading. Our approach combines vision-based perceptual processing with top-down reasoning to map low-level observable behaviour into high-level mental states.

3.1 Representation of Mental States

As shown in Fig. 1, we use Dynamic Bayesian Networks (DBNs) [14] to model the unfolding of mental states over time $P(\mathbf{X}[t])$, where $\mathbf{X}$ is a vector of events corresponding to different mental states. A DBN is a graph that represents the

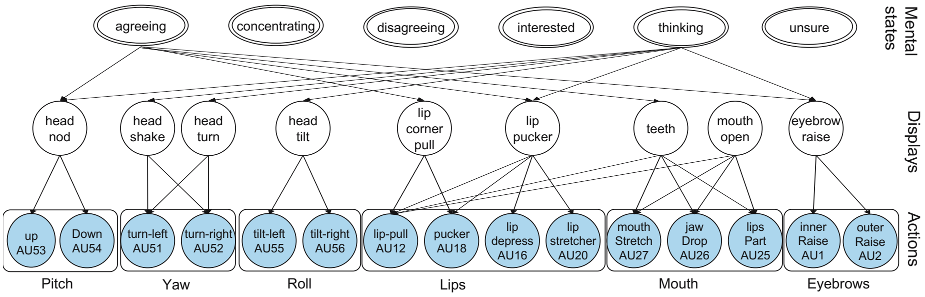

Fig. 1. Multi-level computational model of mind-reading. For clarity, the displays for only two mental states are shown.

causal probability and conditional independence relations among events that evolve over time. The hidden state of each DBN represents an event with two possible outcomes: true whenever the user is experiencing a specific mental state, and false otherwise. The observations or evidence nodes represent the recognized head and facial displays $\mathbf{Y}$. The double circle around a mental state node encodes the temporal dependency between that node in consecutive slices of the network, $X_i[t-1]$ and $X_i[t]$. Having a model for each class means that the hidden state of more than one DBN can be true, so that co-occurring mental states can be represented by the system. The DBN parameters and structure are learnt from exemplar videos using maximum likelihood estimation and feature selection [9].

3.2 Observational Evidence: Head and Facial Displays

The observational evidence consists of the head and facial displays that were recognized up to the current time $P(\mathbf{Y}[t])$ such as a head nod or smile. Each display is represented as a Hidden Markov Model (HMM) of a sequence of head/facial actions, recognized non-intrusively, in real time (Fig. 2). The supported actions and displays are shown in Fig. 1. For instance, a nod is an alternating sequence of head-up and head-down actions. The actions are based on the Facial Action Coding System [5]. Head actions are described by the magnitude and direction of 3 Euler angles, while facial actions are extracted using motion, shape and colour analysis of the lips, mouth and eyebrows. Details of head/facial action recognition and HMM topologies can be found in el Kaliouby [8].

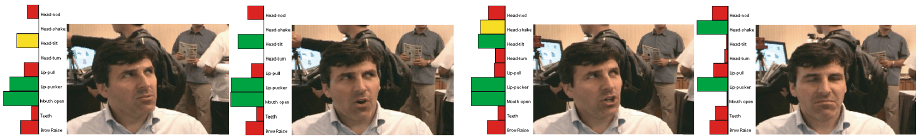

Fig. 2. Real time display recognition (frames sampled every 0.7s). The bars represent the output probabilities of the HMM classifiers (top to bottom): head nod, shake, tilt, turn, lip corner pull, lip pucker, mouth open, teeth and eye-brow raise.

3.3 Inference Framework

Inference involves recursively updating the belief state of hidden states based upon the knowledge captured in the DBNs and available evidence—the head and facial displays that are recognized throughout a video, their dynamics (duration, relationship to each other, and when in the video they occur) and previous mental state inferences. We implement the inference framework as a sliding window of evidence (Algorithm 1). At any instant t, the observation vector that is input to the inference engine is a vector of the w most-recent displays $\mathbf{Y}[t-w:t]$, and the corresponding most-recent mental state inferences $P(\mathbf{X}[t-w:t-1])$. The output is a probability that the observation vector was generated by each of the DBNs. The inference engine uses the unrolled-junction-tree algorithm [14].

Algorithm 1. Mental state inference

```
Objective: $P(X_i[t])$, the belief state of $1 \leq i \leq x$ mental states over time $1 \leq t \leq T$
Given: $x$ DBNs with $y$ observations nodes; evidence length $w$ and sliding factor $dw$
  Instantiate inference engine
  for all $t$ in $w$ time slices do
    Get current observations $\mathbf{Y}[t]$
  for all $t$ in $T$ time slices do
    Enter evidence so far: $\mathbf{Y}[t-w:t]$ and $P(\mathbf{X}[t-w:t-1])$
    Calculate marginal probabilities $P(\mathbf{X}[t])$
    Advance window $t = t + dw$
    Get current observations $\mathbf{Y}[t]$
```

4 Experimental Evaluation

In el Kaliouby and Robinson [9], we trained and tested our system on videos from the Mind-Reading DVD (MR) [2], a guide to emotions developed for Autism Spectrum Disorders. An upper bound of 88.9% and an average accuracy of 77.4% was achieved for *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. To test if our system generalizes beyond the controlled videos in MR, we collected videos at the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2004). Fig. 3 (left) shows frames of both corpora.

4.1 The CVPR 2004 Corpus

We asked 16 conference attendees to act six mental states: *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. The volunteers were *not* given any instructions on *how* to act the mental states, which resulted in considerable within-class variation between the 16 videos of each emotion. They were asked to name the mental state they would act immediately before they started; this was later used to label the videos. Unlike prevalent facial expression databases [10], we placed no restrictions on the head or body movements of volunteers. All 16 volunteers were aged between 16 and 60 and worked in computer-science

Table 1. Characteristics of CVPR corpus “actors”. Gender: Male ● Female ○; Ethnicity: White ● Asian ○; Glasses ● Facial hair ○; Looking down ● Talking ○.

Subject ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Gender	●	●	○	●	●	●	●	●	○	●	○	●	●	●	●	●
Ethnicity	●	●	●	●	●	○	●	●	●	●	●	●	●	●	●	○
Glasses/Facial hair	○				●	●					● ○					
Frontal	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●
Looking down/Talking		○	○	●				○	●	○						

or engineering; most were males of a white ethnic origin[2]. Their characteristics are summarized in Table 1. The videos were captured at 30 fps at a resolution of 320x240 and were labelled using the audio accompanying the footage. The background of the videos is dynamic: people were moving in and out of the neighbouring demonstration booth. We just relied on the lighting in the conference room at the time. The face-size varies within and between videos as the volunteers moved toward/away from the camera. By contrast, the actors in MR had a frontal pose at a constant distance from the camera, none wore glasses or had facial hair and the videos all had a uniform white background and the lighting was professionally set up. Eight videos (15s) were discarded: three lasted less than two seconds which is when the first DBN invocation occurs, and the system failed to locate the face in five videos. We used the remaining 88 videos (313s).

4.2 Human Baseline

Having been posed by people who are not professional actors, the CVPR videos are likely to include incorrect or bad examples of a mental state, and are weakly labelled. To establish a baseline with which to compare the results of the system, we tested how a panel of people would classify the videos. A forced-choice procedure was adopted, with six choices on each question: *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking*, *unsure*. Chance responding is 16.7%. Participants were shown a video on a projection screen, and then asked to circle only one mental state word that best matched what the person in the video was feeling. The panel consisted of 18 participants (50.0% male, 50.0% female), mostly software developers between the ages of 19 and 28[3]. The test generated 88 trials per participant for a total of 1584 responses. The distribution of results is shown in Fig. 3 (right). The percentage of correct answers range from 31.8% to 63.6% (mean=53.03%, SD=0.068). The agreement-score of a video—the percentage of panel participants who assigned the same label to a video—varied between 0-100%. Only 11% of the videos achieved an agreement-score of 85% or more on the truth label of the video; these were deemed as good examples of mental states. The confusion matrix of responses is shown in Fig. 5 (left). The classification rate is highest for *disagreeing* (77.5%) and lowest for *thinking* (40.1%). For a false positive rate of 9.4%, the recognition accuracy of the panel was 54.5%.

[2] Ethnicity defined as in the latest UK census, The Focus on Ethnicity and Identity.

[3] The panel did not know any of the “actors” on the CVPR corpus.

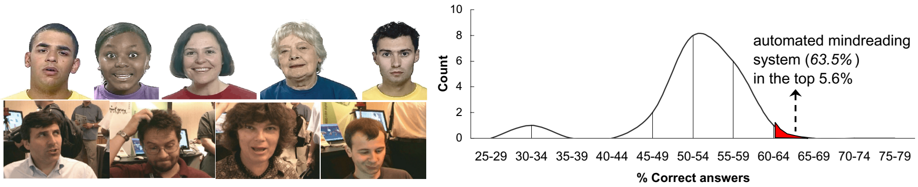

Fig. 3. (top-left) Mind Reading DVD; (bottom-left) CVPR corpus; (right) Distribution of human responses. The accuracy of our system is also shown.

4.3 Results of Computational Model of Mind-Reading

We trained the system on MR videos and tested it on the 88 videos of the CVPR corpus. A classification is correct if the mental state scoring the minimum error (i.e. largest area under the curve) matches the ground-truth label of the video. Fig. 4 shows an example of a 4.3-second long video labelled as *thinking* (77.8% agreement-score). A (false) head shake, a head tilt, a head turn and a lip-pull were recognized. Since *thinking* spans the largest area and this matches the ground-truth label of the video, this is a correct classification. The results are summarized in Fig. 5 (right). The classification rate is highest for *disagreeing* (85.7%) and lowest for *thinking* (26.7%)—all higher than chance responding (16.7%). For a mean false positive rate of 7.3%, the overall accuracy of the system is 63.5%. Compared with the results of humans classifying the exact set

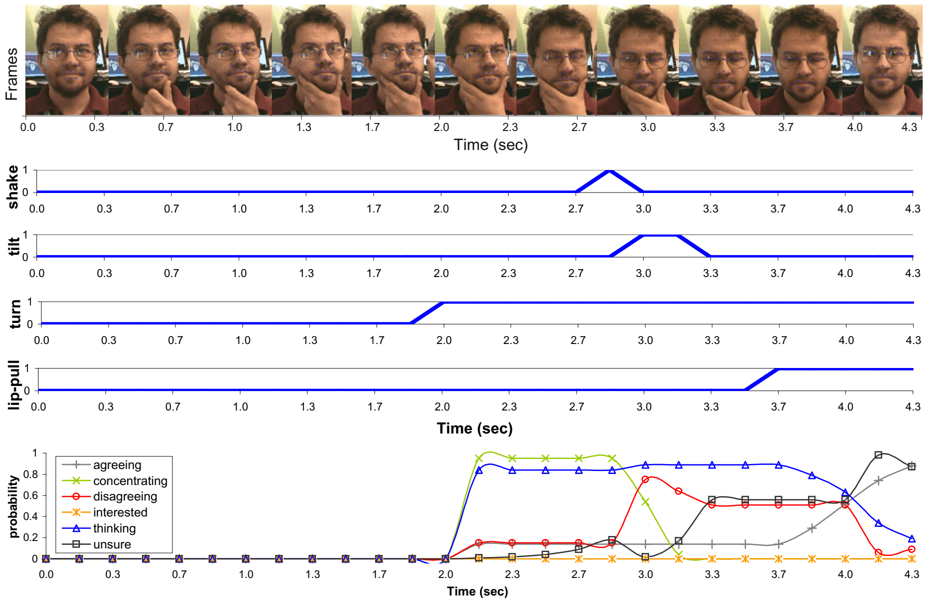

Fig. 4. Mental state inference: (top) frames from a video labelled as *thinking* (CVPR Corpus); (middle) head and facial displays; (bottom) mental state inferences

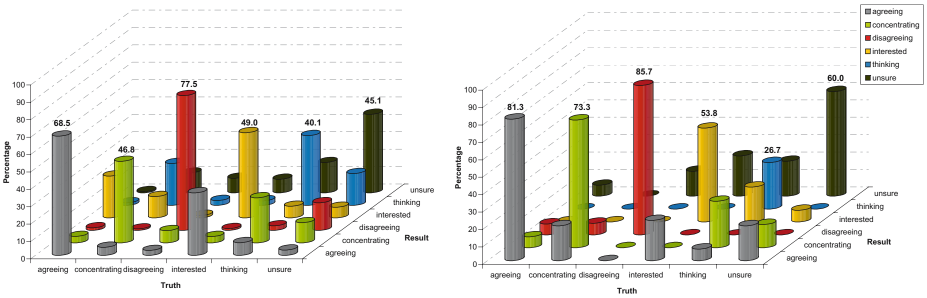

Fig. 5. Confusion matrix of results shown as a 3D bar chart: (left) human recognition results of the CVPR corpus; (right) the system's recognition the CVPR corpus

of videos, the automated mind-reading system scores among the top 5.6% of humans, and 10.2% better than the mean accuracy reported in the sample of 18 people. The result is superimposed on the distribution of human responses shown in Fig. 3 (right). The principal reason why both human recognition (54.5%) and the system's accuracy (63.5%) is generally low is the untrained acting and weak labelling of the CVPR corpus videos. In addition, the recording conditions of the CVPR corpus were much less controlled than that of MR, resulting in challenges in processing these videos automatically (e.g., speech and changes in lighting conditions). The system's recognition accuracy increases to 80% for the 11% of videos with agreement-score of 85% or more, a result similar to that obtained from evaluating the system on MR.

5 Conclusion and Future Directions

This paper described a computational model of mind-reading that infers complex mental states from facial expressions and head gestures in real-time video. The system generalizes well (compared with human recognition) to new examples of mental state enactments, which are posed (and labelled) by lay people in an uncontrolled setup. Albeit posed, the videos were challenging from a machine vision perspective: 1) they contained natural rigid head motion; 2) had overlapping facial expressions and 3) the "actors" expressed the same mental state through different facial expressions, intensities and durations. By recognizing mental states beyond the basic emotions, we widen the scope of applications in which facial analysis systems can be integrated. Moving forward, we will test our system on natural expressions of mental states.

Acknowledgements

Special thanks are due to everyone who took part in our experiments, especially those who volunteered with acting at CVPR 2004. This research was generously funded by the Computer Laboratory's Wiseman Fund, the Overseas Research Student Award, the Cambridge Overseas Trust and Newnham College.

References

1. S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press, 1995.
2. S. Baron-Cohen, O. Golan, S. Wheelwright, and J. J. Hill. *Mind Reading: The Interactive Guide to Emotions*. London: Jessica Kingsley Publishers, 2004.
3. S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. The Reading the Mind in the Eyes Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, 42(2):241–251, 2001.
4. P. Ekman and W. V. Friesen. *Pictures of Facial Affect*. Consulting Psychologists, 1976.
5. P. Ekman and W. V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists, 1978.
6. B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36:259–275, 2003.
7. H. Gu and Q. 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.
8. R. el Kaliouby. *Mind-Reading Machines: Automated Inference of Complex Mental States*. Phd thesis, University of Cambridge, Computer Laboratory, (2005).
9. R. el Kaliouby and P. Robinson. *Real-Time Vision for Human Computer Interaction*, chapter Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures, pages 181–200. Springer-Verlag, 2005.
10. T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive Database for Facial Expression Analysis. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.
11. A. Kapoor, R. W. Picard, and Y. Ivanov. Probabilistic Combination of Multiple Modalities to Detect Interest. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 3, pages 969–972, 2004.
12. G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. R. Movellan. Dynamics of Facial Expression Extracted Automatically from Video. In *Face Processing in Video Workshop at the CVPR2004*, 2004.
13. P. Michel and R. 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.
14. K. P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Phd thesis, UC Berkeley, Computer Science Division, 2002.
15. M. Pantic and L. J. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22:1424–1445, 2000.
16. M. Pardàs, A. Bonafonte, and J. L. Landabaso. Emotion Recognition based on MPEG4 Facial Animation Parameters. In *Proceedings of Internatoinal Conference on Acoustics, Speech and Signal Procssing*, volume 4, pages 3624–3627, 2002.
17. Y.-L. Tian, L. Brown, A. Hampapur, S. Pankanti, A. W. Senior, and R. M. Bolle. Real World Real-time Automatic Recognition of Facial Expressions. *IEEE workshop on Performance Evaluation of Tracking and Surveillance*, 2003.