Mind Reading Machines: Automated Inference of Cognitive Mental States from Video

R. El Kaliouby and P. Robinson
Computer Laboratory
University of Cambridge UK

Abstract

Mind reading encompasses our ability to attribute mental states to others, and is essential for operating in a complex social environment. The goal in building mind reading machines is to enable computer technologies to understand and react to people's emotions and mental states. This paper describes a system for the automated inference of cognitive mental states from observed facial expressions and head gestures in video. The system is based on a multi-level dynamic Bayesian network classifier which models cognitive mental states as a number of interacting facial and head displays. Experimental results yield an average recognition rate of 87.4% for 6 mental states groups: agreement, concentrating, disagreement, interested, thinking and unsure. Real time performance, unobtrusiveness and lack of preprocessing make our system particularly suitable for user-independent human computer interaction.

1. Introduction

People mind read or attribute mental states to others all the time, effortlessly, and mostly subconsciously. Mind reading allows us to make sense of other people's behavior, predict what they might do next, and how they might feel. While subtle and somewhat elusive, the ability to mind read is essential to the social functions we take for granted. A lack of or impairment in mind reading abilities are thought to be the primary inhibitor of emotion and social understanding in people diagnosed with autism (e.g. Baron-Cohen *et. al* [2]).

People employ a variety of nonverbal communication cues to infer underlying mental states, including voice, posture and the face. The human face in particular provides one of the most powerful, versatile and natural means of communicating a wide array of mental states. One subset comprises cognitive mental states such as *thinking*, *deciding* and *confused*, which involve both an affective and intellectual component [4]. Cognitive mental states play an important role in interpreting and predicting the actions of others [22] and as shown in Rozin and Cohen [19] these non-basic mental states occur more often in day to day interactions than the prototypic basic ones (happiness, sadness, anger, fear, surprise and disgust). Because of their intellectual component, cognitive mental states are especially relevant in human computer interaction which often involves problem-solving and decision-making.

Paradoxically, despite the crucial role of cognitive mental states in making sense of people's behaviour, 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 recognizing the set of basic emotions [5, 6, 7, 8, 9, 17, 18, 21].

The recognition of cognitive mental states involves the analysis of multiple asynchronous information sources such as purposeful head gestures, eye-gaze direction, in addition to facial actions [2]. Also, cognitive mental states are only reliably discerned by analysing the temporal dependencies across consecutive facial and head displays [14]. In other words, modelling cognitive mental states involves multi-level temporal abstractions: at the highest level, mental states typically last between 6-8 sec [3]. Displays can last up to 2 sec, while at the lowest level, action units last tenths of seconds.

This paper describes a system for inferring cognitive mental states from video of facial expressions and head gestures in real time. Being unobtrusiveness and fully automated makes the system particularly suitable for user-independent man-machine contexts. To our knowledge, this work makes the first attempt at classifying cognitive mental states automatically.

2. Overview

Our approach combines machine vision and supervised statistical machine learning to model hidden mental states of a person based upon the observable facial and head displays of that person. An overview of the automated mind

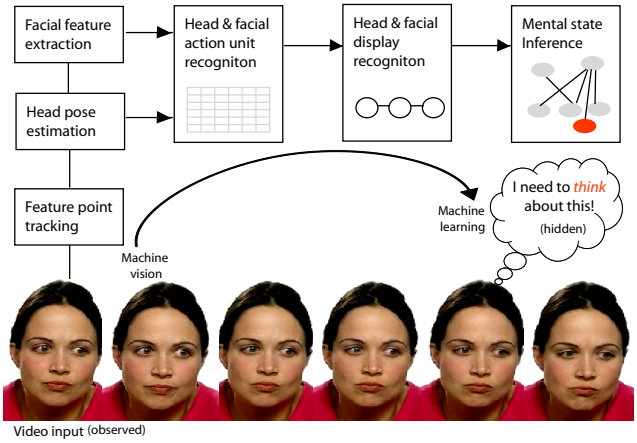


Figure 1: Block diagram of the automated mind reading system

reading system is shown in Figure 1. Video of the face is recorded at 29 frames per second and input to the system in real time. We assume a full frontal view of the face, but take into account variations in head pose and framing inherent in video-based interaction. The vision-based component recognizes dynamic head and facial displays from video. It locates and tracks fiducial landmarks across an image, then estimates head pose from expression-invariant feature points. The head pose parameters depict head action units. Facial feature motion, shape and color descriptors identify facial action units. Head and facial actions are combined temporally in a hidden Markov model (HMM) framework to recognize displays.

The inference component makes use of dynamic graphical models, specifically dynamic Bayesian networks (DBNs) that represent high-level cognitive mental states given observed displays. A separate model of each mental state is learned allowing the system to be in more than one mental state at a time. This is particularly useful for modelling mental states that are not mutually exclusive. The use of DBNs makes it possible to later add eye-gaze and context to map multiple information sources to mental states. By exploiting the different temporal scale of each level the mind reading system runs in real time. For example, instead of invoking a mental state inference on every frame, approximately 20 inferences are made in a video 6 seconds long (190 frames). In addition, each level of the system is implemented as a sliding window to make it possible to run the system for an indefinite duration.

3 Head and facial action unit analysis

Twenty four facial landmarks are detected using a face template in the initial frame, and their positions tracked across the video. The system builds on Facestation [1], a feature point tracker that supports both real time and offline tracking of facial features on a live or recorded video stream. The tracker represents faces as face bunch graphs [23] or stack-like structures which efficiently combine graphs of individual faces that vary in factors such as pose, glasses, or physiognomy. The tracker outputs the position of twenty four feature points, which we then use for head pose estimation and facial feature extraction.

3.1 Extracting head action units

Natural human head motion typically ranges between 70-90^o of downward pitch, 55^o of upward pitch, 70^o of yaw (turn), and 55^o of roll (tilt), and usually occurs as a combination of all three rotations [16]. The output positions of the localized feature points are sufficiently accurate to permit the use of efficient, image-based head pose estimation. Expression invariant points such as the nose tip, root, nostrils, inner and outer eye corners are used to estimate the pose. Head yaw is given by the ratio of left to right eye widths. A head roll is given by the orientation angle of the two inner eye corners. The computation of both head yaw and roll is invariant to scale variations that arise from moving toward or away from the camera. Head pitch is determined from the vertical displacement of the nose tip normalized against the distance between the two eye corners to account for scale variations. The system supports up to 50^o, 30^o and 50^o of yaw, roll and pitch respectively. Pose estimates across consecutive frames are then used to identify head action units. For example, a pitch of 20^o degrees at time t followed by 15^o at time $t+1$ indicates a downward head action, which is AU54 in the FACS coding [10].

3.2 Extracting facial action units

Facial actions are identified from component-based facial features (e.g. mouth) comprised of motion, shape and colour descriptors. Motion and shape-based analysis 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. Color-based analysis is computationally efficient, and is invariant to the scale or viewpoint of the face, especially when combined with feature localization (i.e. limited to regions already defined by feature point tracking).

The shape descriptors are first stabilized against rigid head motion. For that, 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, a 2D projection of the approximated real point around which

the head rotates in 3D space. 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.

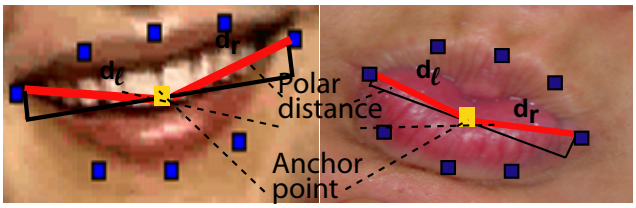


Figure 2: Polar distance in determining a lip corner pull and lip pucker

On each frame, the polar distance between each of the two mouth corners and the anchor point is computed. The average percentage change in polar distance calculated with respect to an initial frame is used to discern mouth displays. An increase or decrease of 10% or more, determined empirically, depicts a lip pull or lip pucker respectively (Figure 2). In addition, depending on the sign of the change we can tell whether the display is in its onset, apex, offset. The advantages of using polar distances over geometric mouth width and height (which is what is used in Tian *et al.* [20]) are support for head motion and resilience to inaccurate feature point tracking, especially with respect to lower lip points.

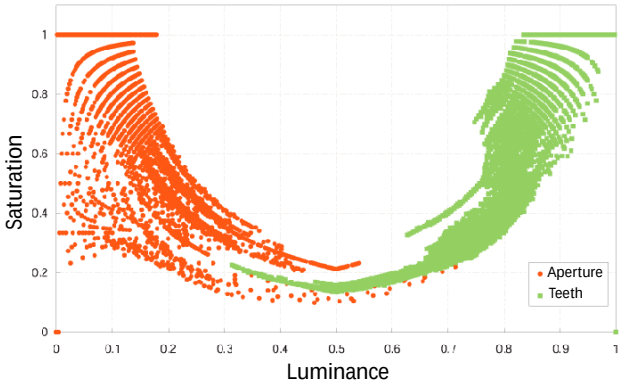


Figure 3: Plot of aperture (red) and teeth (green) in luminance-saturation space

The mouth has two color regions that are of interest: aperture and teeth. The extent of aperture present inside the mouth depicts whether the mouth is closed, lips parted, or jaw dropped, while the presence of teeth indicates a mouth stretch. Figure 3 shows a plot of teeth and aperture samples in luminance-saturation space. Luminance, given by the relative lightness or darkness of the color, acts as a good discriminator for the two types of mouth regions. A sample of $n = 125000$ pixels was used to learn the probability distribution functions of aperture and teeth. A lookup table defining the probability of a pixel being aperture given its luminance is computed for the range of possible luminance values (0% for black to 100% for white). A similar lookup table is computed for teeth. Online classification into mouth actions proceeds as follows: For every frame in the sequence, we compute the luminance value of each pixel in the mouth polygon. The luminance value is then looked up to determine the probability of the pixel being aperture or teeth. Depending on empirically determined thresholds the pixel is classified as aperture or teeth or neither. Finally, the total number of teeth and aperture pixels are used to classify the mouth region into closed (or lips part), jaw drop, or mouth stretch. Figure 4 shows classification results of 1312 frames into closed, jaw drop and mouth stretch.

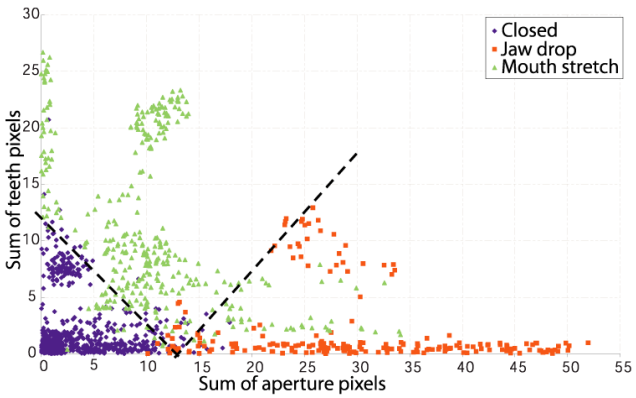


Figure 4: Classifying 1312 mouth regions into closed, jaw drop or stretch

4 Head and facial display recognition

Facial and head actions are quantized and input into left-to-right HMM classifiers to identify facial expressions and head gestures. Each is modelled as a temporal sequence of action units (e.g. a head nod is a series of alternating up and down movement of the head). In contrast to static classifiers which classify single frames into an emotion class, HMMs model dynamic systems spatio-temporally, and deal with the time warping problem. In addition, the convergence of recognition computation may run in real time, a desirable aspect in automated facial expression recognition systems for human computer interaction.

We devise several HMM topologies for the recognition of the displays. For instance the head nod HMM is a 4-state, 3 symbol HMM, where the symbols correspond to head up, head down, and no action. We use a similar

topology for head shakes and supported mouth displays. For tilt and turn displays we use a 2-state HMM with 7 observable symbols. The symbols encode the intensity of the tilt and turn motions. Maximum likelihood training is used to determine the parameters of each HMM model $\lambda = \{\Lambda, \beta, \pi\}$ offline, described by transition probabilities, the probability distributions of the states, and priors.

For each model λ and a sequence of observations $O = \{o_1, o_2, ..., o_T\}$ forward-backward algorithm determines the probability that the observations are generated by the model. Forward-backward is linear in T, so is suitable for running in real time.

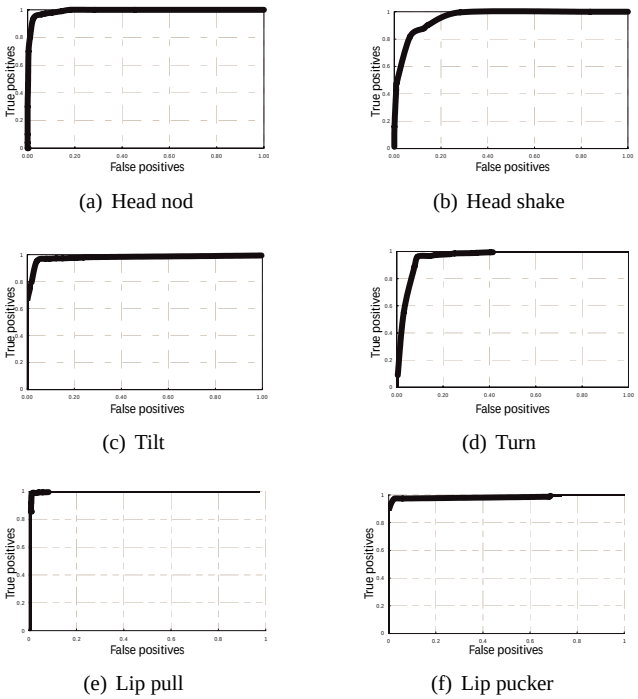
5 Cognitive mental state inference

The HMM level outputs a likelihood for each of the facial expressions and head displays. However, on their own, each display is a weak classifier that does not entirely capture an underlying cognitive mental state. Bayesian networks have successfully been used as an ensemble of classifiers, where the combined classifier performs much better than any individual one in the set [15]. In such probabilistic graphical models, hidden states (the cognitive mental states in our case) influence a number of observation nodes, which describe the observed facial and head displays. In dynamic Bayesian networks (DBN), temporal dependency across previous states is also encoded.

Training the DBN model entails determining the parameters and structure of a DBN model from data. Maximum likelihood estimates is used to learn the parameters, while sequential backward elimination picks the (locally) optimal network structure for each mental state model. More details on how the parameters and structure are learnt can be found in [13].

6 Experimental evaluation

For our experimental evaluation we use the Mind reading dataset (MR) [3]. MR is a computer-based guide to emotions primarily collected to help individuals diagnosed with Autism recognize facial expressions of emotion. A total of 117 videos, recorded at 30 fps with durations varying between 5 to 8 seconds, were picked for testing. The videos conveyed the following cognitive mental states: *agreement*, *concentrating*, *disagreement*, *thinking* and *unsure* and *interested*. There are no restrictions on the head or body movement of actors in the video. The process of labelling involved a panel of 10 judges who were asked could this be the emotion name? When 8 out of 10 agree, a statistically significant majority, the video is included in MR. To our knowledge MR is the only available, labelled



(a) Head nod (b) Head shake

(c) Tilt (d) Turn

(e) Lip pull (f) Lip pucker

Figure 5: ROC curves for head and facial displays

resource with such a rich collection of mental states and emotions, even if they are posed.

We first evaluate the classification rate of the display recognition layer and then the overall classification ability of the system.

6.1 Display recognition

We evaluate the classification rate of the display recognition component of the system on the following 6 displays: 4 head displays (head nod, head shake, tilt display, turn display) and 2 facial displays (lip pull, lip pucker). The classification results for each of the displays are shown using the Receiver Operator Characteristic (ROC) curves (Figure 5). ROC curves depict the relationship between the rate of correct classifications and number of false positives (FP). The classification rate of display d is computed as the ratio of correct detections to that of all occurrences of d in the sampled videos. The FP rate for d is given by the ratio of samples falsely classified as d to that of all non-d occurrences. Table 2 shows the classification rate that the system uses, and the respective FP rate for each display.

A non-neutral initial frame is the main reason behind undetected and falsely detected displays. To illustrate this, consider a sequence that starts as a lip pucker. If the lip pucker persists (i.e. no change in polar distance) the pucker display will pass undetected. If on the other hand, the pucker returns to neutral (i.e. increase in polar distance)

it will be falsely classified as a lip pull display. This problem could be solved by using the polar angle and color analysis to approximate the initial mouth state. The other reason accounting for misclassified mouth displays is that of inconsistent illumination. Possible solutions to dealing with illumination changes include extending the color-based analysis to account for overall brightness changes or having different models for each possible lighting condition.

6.2 Mental state recognition

We then evaluate the overall system by testing the inference of cognitive mental states, using leave-5-out cross validation. Figure 6 shows the results of the various stages of the mind reading system for a video portraying the mental state *choosing*, which belongs to the mental state group *thinking*. The mental state with the maximum likelihood over the entire video (in this case *thinking*) is taken as the classification of the system.

87.4% of the videos were correctly classified. The recognition rate of a mental class m is given by the total number of videos of that class whose most likely class (summed over the entire video) matched the label of the class m. The false positive rate for class m (given by the percentage of files misclassified as m) was highest for *agreement* (5.4%) and lowest for *thinking* (0%). Table 2 summarizes the results of recognition and false positive rates for 6 mental states.

A closer look at the results reveals a number of interesting points. First, onset frames of a video occasionally portray a different mental state than that of the peak. For example, the onset of *disapproving* videos were (mis)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 with the person being *unsure*. Second, subclasses that do not clearly exhibit the class signature are easily misclassified. For example, the *assertive* and *decided* videos in the *agreement* group were misclassified as *concentrating*, as they exhibit no smiles, and only very weak head nods. 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.

7 Applications and conclusion

The principle contribution of this paper is a multi-level DBN classifier for inferring cognitive mental states from videos of facial expressions and head gestures in real time. The strengths of the system include being fully automated, user-independent, and supporting purposeful head displays while de-coupling that from facial display recognition. We reported promising results for 6 cognitive mental states on a medium-sized posed dataset of labelled videos. Our current research directions include:

1. testing the generalization power of the system by evaluating a larger and more natural dataset
2. exploring the within-class and between-class variation between the various mental state classes, perhaps by utilizing cluster analysis and/or unsupervised classification
3. adding more mental state models such as *comprehending*, *bored* and *tired*, which like the ones already reported in this paper are relevant in an HCI context.

On the applications front we are working on integrating the system with instant messaging [12] to add spontaneity of interaction. In addition, we are building a prototype of an "emotional hearing aid", an assistive tool for people diagnosed with Asperger's Syndrome [11] designed to provide advice on emotion understanding from video. We believe that the work presented is an important step towards building mind reading machines.

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. 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] Facestation tracking cti, 2002.

[2] 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.

[3] S. Baron-Cohen and T. H. E. Tead. Mind reading: The interactive guide to emotion, 2003.

[4] 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.

[5] M. Bartlett, G. Littlewort, B. Braathen, and T. Sejnowski. A prototype for automatic recognition of spontaneous facial actions. In *Advances in Neural Information Processing Systems*, volume 15, 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

Table 1: Recognition results for head and facial displays from the Mind reading dataset

Display	#Train.	#Displays	Class. Rate(%)	#non-Displays	FP rate(%)
Head nod	37	256	96.7	2637	5.96
Head shake	30	231	87.88	2392	12.75
Head Tilt	93	953	95.0	1648	4.0
Head Turn	126	1219	96.14	1397	8.95
Lip pull	20	427	98.83	2196	2.73
Pucker	49	156	97.44	2467	1.113

Table 2: Recognition results for mental state inference from the Mind reading dataset

Mental state	#Videos	Class. Rate(%)	#non-Displays	FP rate(%)
Agreement	25	88.1	81	5.4
Concentrating	10	90	96	2.08
Disagreement	13	80	103	0.97
Interested	10	90	96	1.04
Thinking	18	88.9	88	0
Unsure	30	90	76	5.2

Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, volume 1, pages 595–604, 2003.

[7] J. Cohn. *What the face reveals(2nd edition)*, chapter Automated analysis of the configuration and timing of facial expression. Oxford University Press Series in Affective Science. New York: Oxford, 2004.

[8] 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.

[9] 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.

[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 and P. Robinson. The emotional hearing aid: an assistive tool for children with asperger's syndrome. In *International Workshop on Universal Access and Assistive Technology*, pages 244–246, 2004.

[12] R. el Kaliouby and P. Robinson. Faim: Integrating automated facial affect analysis in instant messaging. In *Proceedings of ACM Intelligent User Interfaces Conference*, pages 244–246, 2004.

[13] R. el Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *International Workshop on Real-Time Computer Vision for Human Computer Interaction*, 2004.

[14] 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.

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

[16] T. Kurz. *Stretching Scientifically: A Guide to Flexibility Training*. Stadion Publishing Co, 2003.

[17] J. Lien, A. Zlochower, J. Cohn, and T. Kanade. Automated facial expression recognition. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.

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

[19] 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.

[20] Y.-L. Tian, T. Kanade, and J. Cohn. Robust lip tracking by combining shape, color and motion. In *Proceedings of the 4th Asian Conference on Computer Vision*, January 2000.

[21] 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.

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

[23] L. Wiskott, J. Fellous, N. Krger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

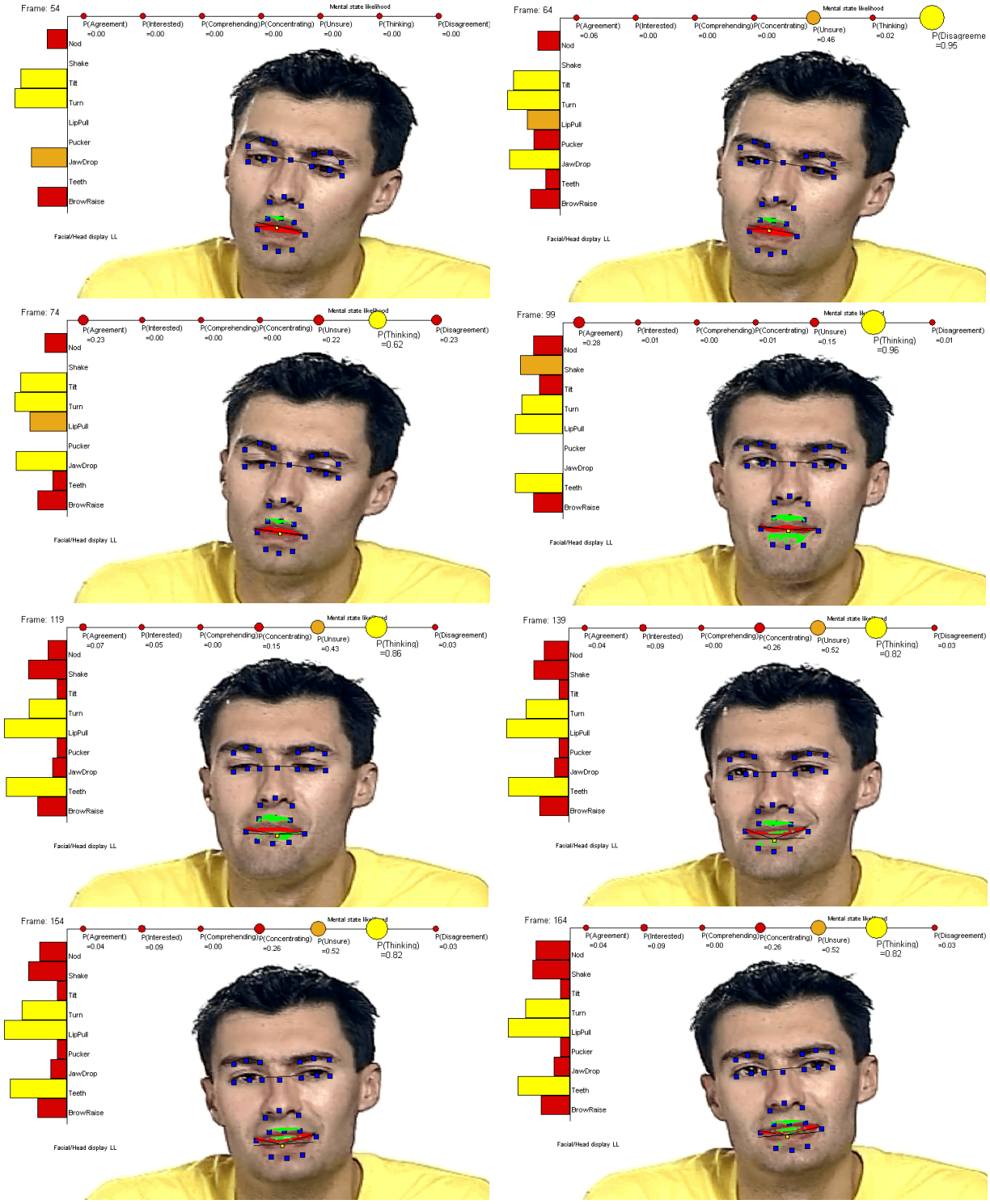


Figure 6: The status of the dynamic Bayesian networks for 7 mental states shown at 8 inference instances in a video of *choosing*, which belongs to the *thinking* group. The vertical axis encodes the output of the facial/head display HMM classifiers. Longer bars (also color coded in yellow) represent a higher likelihood of a display. Displays from top to bottom are: nod, shake, tilt, turn, lip corner pull, lip pucker, jaw drop, mouth stretch (teeth), and eye brow raise. The horizontal axis encodes the likelihood for 7 mental states. Larger circles (shown in yellow) encode higher likelihood of a mental state. Mental states from left to right are: *agreement*, *interested*, *comprehending*, *concentrating*, *unsure*, *thinking* and *disagreement*. For the first instance, the likelihoods of all mental states are 0 (indicated by the small red circles). As the video progresses, the likelihoods change. The mental state with the maximum likelihood over the entire video (in this case *thinking*) is taken as the classification of the system.