International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 1, Issue 5 (Nov-Dec 2013), PP. 88-93

MIND READING COMPUTER MACHINE

Vishal Bhalla, Tarun Yadav, Vishal Sharma, Sumit Kumar Dass
Computer science engineering
Dronacharya college of engineering Gurgaon.
Gurgaon(India).

***Abstract*—This 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 multilevel 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, and disagreement, interested, thinking and unsure. Real time performance, unobtrusiveness and lack of preprocessing make our system particularly suitable for User-independent human computer interaction.**

I. INTRODUCTION

As you were reading through the abstract, your mind engaged in a series of calculations designed to figure out what I meant in using the specific words and phrases that I used in writing it. Your capacity to "mind read" attempted to identify and ascribe to me a coherent set of beliefs, intentions, desires and other mental states that might have led me to write what I wrote. This powerful compulsion to predict and explain the behavior of other agents is not only active while reading text, but also while engaged in dialogue, while observing everyday human action, while competing or cooperating, when engaged with fiction of any type, and possibly when planning for the future or learning from past episodes. The central cognitive activity involved in mindreading is the ascription of mental states from one agent to another. If Max observes Sally walking to the kitchen, he might infer that Sally is hungry, wants something to eat and will walk to the refrigerator because she thinks there is food inside. Max ascribes a number of mental states to Sally including her belief that food is in the fridge, that she desires to eat, and that she intends to walk to the fridge in order to get a snack. However, he likely does not ascribe other less relevant but logically possible mental states, such as Sally wanting to get something from the refrigerator and her believing that 89 is in the set of prime numbers. Although it seems odd to consider the latter as an example, such inferences are not only warranted but demanded on certain formal accounts of reasoning about beliefs. 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*). 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. Cognitive mental states play an important role in interpreting and predicting the actions of others and as shown in Rosin and Cohen 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 behavior 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) , or only go as far as recognizing the set of basic Emotions. 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. Also, cognitive mental states are only reliably discerned by analyzing the temporal dependencies across consecutive facial and head displays. In other words, modeling cognitive mental states involves multilevel temporal abstractions: at the highest level, mental states typically last between 6-8 sec. 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.

1.1 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 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. 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.

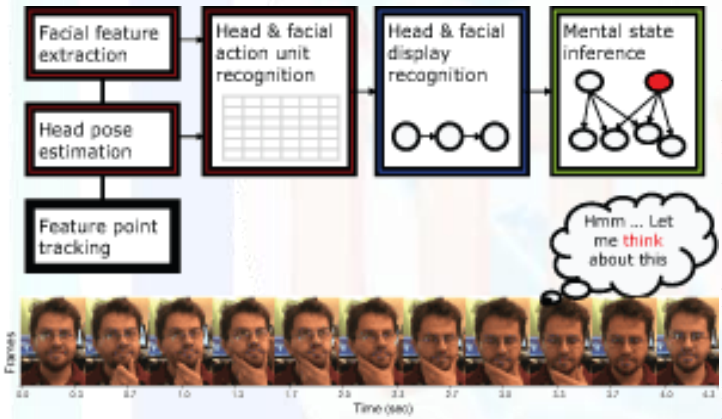

Figure1

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 fiducially 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 modeling 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

1.2 Methodology and Modeling

In a recent Cognitive Science article, Casemates, Langley and Bello (2008) argued for three core criteria to be applied in the evaluation of models for higher-order cognition. These three criteria are ability, breadth, and parsimony. Generally speaking, by ability we meant the general capacity of a model to account for human-level competence with respect to the phenomena under investigation. By breadth, we meant that the model is capable of accounting for a variety (if not the preponderance) of phenomena-related results, including capturing competence-related trends across a sufficiently large space of human data. By parsimony, we meant that the model displays both ability and sufficient breadth without multiplying cognitive mechanisms (or representations) beyond the demands imposed by our most current data. As I shift discussion toward existing computational approaches to mindreading, I will argue that typically employed assumptions in both AI and computational cognitive science fail on at least one of these criteria. As a matter of methodology, I am committed to not only giving a computational explanation of Mind reading as a capacity, but also providing hypotheses for how it might be degraded or even fail outright. The strategy I adopt is to assume that error-prone mindreading is the result of cognitive systems that evolved for purposes other than mindreading and have since been re-purposed to the task of understanding other minds. One might argue that building a cognitive system that is prone to attribution errors seems wasteful or otherwise silly. I think that this remains to be seen. There are many types of social interaction where one agent benefits by having the ability to reason about the kinds of attribution errors made by another agent. For example, games like poker would be much less interesting for expert players if they were not able to apply a fairly rich model of errors to their advantage, even if they have no consciously accessible theory of attribution errors to draw from. The semantics of important social concepts like stereotyping would be difficult to capture in.

II. RELATED WORK

.

2.1 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 Face station, 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 bunk graphs 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.

2.2 Extracting head action units

Natural human head motion typically ranges between 70- 90o of downward pitch, 55o of upward pitch, 70o of yaw (turn), and 55o of roll (tilt), and usually occurs as a combination of all three rotations . 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 50o, 30o and 50o of yaw, roll and pitch respectively. Pose estimates across consecutive frames are then used to identify head action units. For example, a pitch of 20o degrees at time t followed by 15o at time t + 1 indicates a downward head action, which is AU54 in the FACS coding.

2.3 Extracting facial action units

Facial actions are identified from component-based facial features (e.g. mouth) comprised of motion, shape and color 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 (Xp, Yp) 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.

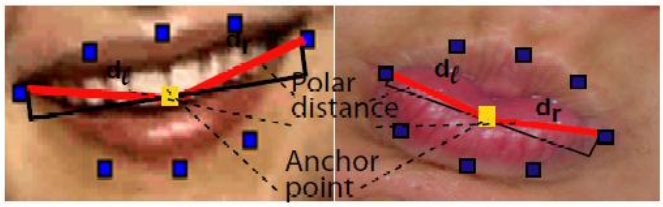

Figure 2: Polar distance in determining a lip corner pull and lip pucker

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.

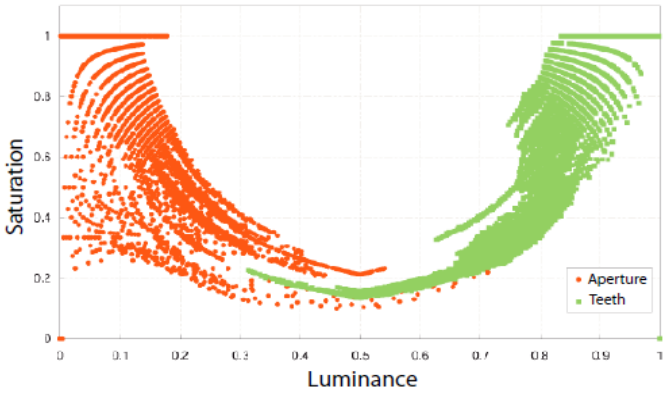

Figure 3: Plot of aperture (red) and teeth (green) in luminance-saturation space

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

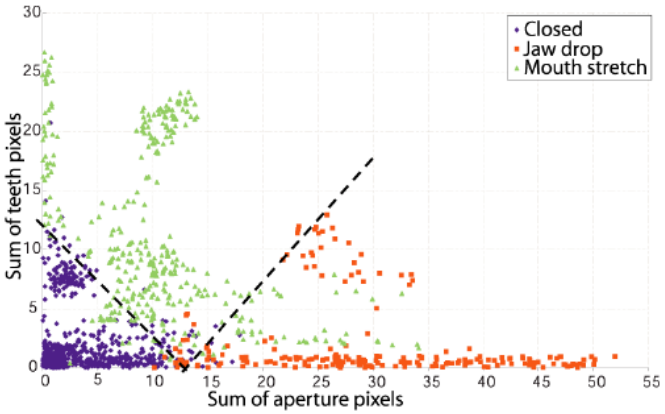

Figure 4: Classifying 1312 mouth regions into closed, jaw drop or stretch

2.4 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 modeled 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 spatial-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 _ = {_, _, _} offline, described by transition probabilities, the probability distributions of the states, and priors. For each model _ and a sequence of observations O = {o1, o2... out } 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.

2.5 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 he 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 was (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.

III. Conclusion

Mindreading represents one of the most complicated and interesting cognitive activities in which we routinely engage. As such, we ought to take it seriously as a major desideratum in the development of cognitive systems. The overall aim of this paper has been to illustrate the complexities of mindreading and the relative difficulty in trying to account for them using assumptions that typify standard techniques in AI. Many of these assumptions are prescriptive by their nature, and enforce constraints on rationality that are rarely satisfied during real-world episodes of mindreading or even during controlled studies performed in laboratory settings. I have argued that a deflationary account of the mental states of others consisting primarily of counterfactual simulations and inheritance explains the close relationship between performance on mindreading tasks and data on entertaining pretenses. Standard accounts of propositional attitudes assume a sharp delineation between mental states, usually related to the kinds of actions that they tend to motivate. At best I think we have seen that this assumption is questionable, and at at worst it seems wrong. When taken to unreasonable extremes, it seems as if totally decoupling mental states from one another at the level of implementation makes it difficult to

explain engagement with fiction, empathy, wishful thinking, self-deception, pretense, delusions or hallucinations. While some theorists see these as unfortunate outliers, I have argued that mindreading-enabled systems should be able to recognize them in others and modify their interaction strategies accordingly. Having a system that initially is capable of exhibiting all of these behaviors and using simulation to recognize them in others seems to be a reasonable alternative to the rather ugly option of trying to axiomatize them in service of reasoning about them. I have further argued that inheritance rules implemented as soft constraints lets us fit a wide swath of data on mindreading than spans the gap between totally incorrect and perfectly correct attributions. Under assumptions of unlimited inferential resources, this range of attributions accounts for systematic mispredictions and perfectly rational epistemic inference alike. There is much work to be done to flesh out my suggestions into a robust implementation. While the representation of inheritance as soft constraints allows for variance in the attribution process, it is unclear how to systematically link costs on constraints to other features of ongoing cognition, including explicit judgments and resource limitations in the cognitive system. I have also intentionally left the discussion of learning new inheritance constraints from successful and unsuccessful episodes of mindreading as an open issue. The issue of whether or not such learning is automatic or intentionally initiated remains open, and computational expressions of the learning process are equally undeveloped. The influence of affect, emotions, feelings, and physiological variables on inheritance is completely unexplored in this paper, as is the question of how to reason when uncertain about the mental states of the target or when knowing the target to be uncertain about a proposition of interest. I have also not spent any time on the relationship between third person mindreading and introspection. In short, this paper has barely scratched the surface, but I hope the suggestions that I have provided will serve as a good starting point for researchers who are interested in accounting for both mindreading competence and architecture-level performance in a parsimonious way.

IV. Acknowledgment

I am indebted to Pat Langley for his careful reviewing of earlier versions of this document and his subsequent suggestions, which have dramatically improved the structure of the paper. I would also like to thank Will Bridewell, Selmer Bringsjord and Bertram Malle for their trenchant commentary on the conceptual contents. Finally, my deepest gratitude goes out to Nick Cassimatis for his continued mentorship and seminal contributions in the development of these ideas.

V. References

[1] Bello, P. (2011). Shared representations of belief and their effects on action selection: A preliminary computational cognitive model. Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society (pp. 2997–3002). Boston, MA.

[2] Bello, P., Bignoli, P., & Cassimatis, N. (2007). Attention and association explain the emergence of reasoning about false belief in young children. Proceedings of the Eighth International Conference on Cognitive Modeling (pp. 169–174). University of Michigan, Ann Arbor, MI.

[3] Bello, P., & Guarini, M. (2010). Introspection and mindreading as mental simulation. Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society (pp. 2022–2028).

[4] Portland, OR.Cassimatis, N. (2006). A cognitive substrate for human-level intelligence. AI Magazine, 27, 45–56.

[5] Cassimatis, N., Bello, P., & Langley, P. (2008). Ability, parsimony and breadth in models of higherordercognition. Cognitive Science, 33, 1304–1322.

[6] Garber, D. (1983). Old evidence and logical omniscience in Bayesian confirmation theory. In

[7] J. Earman (Ed.), Testing scientific theories. Minneapolis, MN: University of Minnesota Press.

[8] Hintikka, J. (1962). Knowledge and belief. An introduction to the logic of the two notions. Ithaca, NY: Cornell University Press.

[9] Kovacs, A., Teglas, E., & Endress, A. (2010). The social sense: Susceptibly to others' beliefs in human infants and adults. Science, 330, 1830–1834.

[10] Langley, P., & Choi, D. (2006). A unified cognitive architecture for physical agents. Proceedings of the Twenty-First National Conference on Artificial Intelligence. Boston, MA: AAAI Press.

[11] Langley, P., Laird, J., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges.

[12] Cognitive Systems Research, 10, 141–160.

[13] Leslie, A., Friedman, O., & German, T. (2004). Core mechanisms in theory of mind. Trends in Cognitive Science, 8, 528–533.

[14] Onishi, K., & Baillargeon, R. (2005). Do 15-month old infants understand false beliefs? Science, 308, 255–258.

[15] Rao, A., & Georgeff, M. (1998). Decision procedures for BDI logics. Journal of Logic and Computation,8, 293–343.

[16] Riggs, K., & Peterson, D. (2000). Counterfactual thinking in pre-school children: mental state and causal inferences. In P. Mitchell and K. Riggs (Eds.), Children's reasoning and the mind (pp.87–99). Hove, UK: Psychology Press.

[17] Savitsky, K., Keysar, B., Epley, N., Carter, T., & Sawnson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. Journal of Experimental Social Psychology, 47, 269–273.

[18] Scally, J., Cassimatis, N., & Uchida, H. (2011). Worlds as a unifying element of knowledge representation.

[19] Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems (pp.280–287). Arlington, VA: AAAI Press.
[20] Sim, K. (1997). Epistemic logic and logical omniscience: A survey. International Journal of Intelligent Systems, 12, 57–81.
[21] Sun, R., & Zhang, X. (2003). Accessibility versus action-centeredness in the representation of cognitive skills. Proceedings of the Fifth International Conference on Cognitive Modeling. Bamberg, Germany.
[22] Tamir, D., & Mitchell, J. (in press). Anchoring and adjustment during social inferences. Journal of Experimental Psychology: General.
[23] Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Structure, statistics, and abstraction. Science, 331, 1279–1285.
[24] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science,185, 1124–1131.
[25] Van der Hoek, W., & Lomuscio, A. (2004). A logic of ignorance. Electronic Notes in Theoretical Computer Science, 85, 1–17.