Automated Classification of Complex Mental States from Head and Facial Displays in Video

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

Abstract

This paper presents a system for the automated inference of complex mental states from observed facial expressions and head gestures in video. Head actions are identified from image-based pose estimation. Facial actions, including asymmetric ones, are extracted from motion, color and shape parameters. A multi-level dynamic Bayesian network classifier models complex mental states as a sequence of interacting facial and head displays. We evaluate the recognition of six complex mental states groups. Our experiments demonstrate the effectiveness of our approach to fully automated, head motion resilient recognition of displays representing complex mental states.

1. Introduction

Over the past decade, significant progress has been made in identifying basic units of muscular activity in the human face (action units (AUs) of the Facial Action Coding System [3]), and classifying them into the set of basic emotions. Emotions, of course, are only one subset of the range of mental states that people experience. The human face communicates a wide array of mental states, including complex ones such as *thinking* and *confused*.

We describe a system for inferring complex mental states from video of facial expressions and head gestures in real time. Commodity hardware is used, such as a commercial digital camcorder placed near the user's monitor and connected to a standard PC. We assume a full frontal view of the face, but take into account variations in head pose and framing inherent in video-based interaction. Videos are captured at a frame rate of 30 per second.

The automated inference of complex mental states from continuous video is a two-step process. The first involves recognising "observed" head gestures and facial expressions as follows: Action units are identified from feature tracking, shape and color descriptors. Our approach is similar to Tian *et al.* [7] but also supports out-of-plane rigid head motion and facial action asymmetry. Consecutive action units are then quantised and combined temporally into sequences, and classified into meaningful displays using hidden Markov model (HMM) classifiers. The second process involves combining these displays in a dynamic Bayesian network (DBN) to infer the underlying, "hidden" complex mental state. 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. Garg et al's [2, 5] fusion of audio and visual cues using DBNs provides the inspiration for our present work. DBNs have also been used in unsupervised learning and clustering of facial displays (Hoey and Little [6]). Real time recognition is achieved by temporally abstracting displays so that a classification per frame is not necessary.

The system makes three principle contributions: 1) explores the characteristics of complex mental states whereas most of the work to date has focused on basic emotions 2) provides an insight into an optimal subset of facial and head displays that are most relevant in identifying the different mental states, which was challenging given the limited expert domain knowledge in this area, and 3) a methodology for inference of complex mental states in real time.

2. Facial and head display recognition

Figure 1 shows the overall structure of the system. A real time facial feature tracker (at 30 fps) locates and tracks 24 facial landmarks from a continuous video stream. Since our goal is to classify head and facial displays simultaneously, the problem of rigid head motion had to be addressed. In-plane and out-of-plane rigid head motion is dealt with by computing image-based pose estimates for each frame. Head action units are identified from pose estimates across consecutive frames. After being stabilised against head motion, facial action units are identified from component-based facial features extracted from motion, color and shape parameters. Based on empirical observations, head and

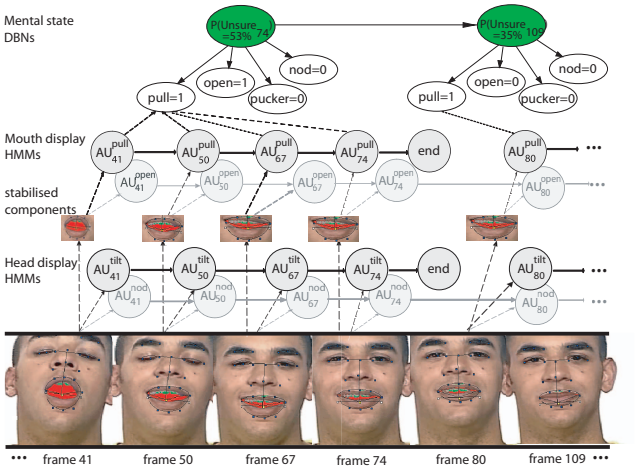

Figure 1: Overview of complex mental state inference. On each frame head action units are identified, the face components are stabilised and then facial action units are extracted from motion, shape and color. HMM classifiers model displays temporally, then DBNs infer underlying mental states given observed displays.

facial actions are detected over 200 millisecond intervals.

Facial and head action units are the building blocks of displays. Consecutive action units are quantised and combined temporally into sequences, typically 6 symbols long, spanning between 0.6 to 1.2 seconds. Sequences are classified into meaningful displays using HMM classifiers. Spatio-temporal modelling of facial and head displays fully exploits the dynamic information inherent in a continuous video stream of the face. To estimate the intensity of a display, parameters such as duration, peak displacement and total energy rate are extracted.

3. Inferring complex mental states

On their own, facial expressions and head gestures are weak classifiers that do not capture the underlying complex 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 [5]. Figure 1 illustrates the DBN model for *unsure*. Each hidden mental state influences a number of observation nodes, which describe the observed facial and head display. In addition, temporal dependency across previous states is also encoded. Maximum likelihood estimates are used to learn the model parameters of 6 mental state groups from data, and the classical forward-backward algorithm is used for inference. Details of feature selection, learning and inference can be found at [4].

4. Experimental evaluation

We sampled 106 videos representing 6 mental state groups: *agreement*, *concentrating*, *disagreement*, *interested*, *thinking* and *unsure* from *Mind Reading*, a computer-based guide to emotions [1]. MR is the only available, labelled resource supporting such a wide range of mental states and emotions, even if they are posed. Approximately 20 inferences are made in a video 6 seconds long, enabling the system to run in real time. The results were promising and showed that some mental states were "closer" to each other and could co-occur.

5. Conclusion

This paper presented a multi-level DBN classifier for inferring complex mental states from videos of facial expressions and head gestures in real time. We reported promising results for 6 complex mental states. The system serves as an important step towards real time integration of facial affect inference in mainstream computing applications.

Acknowledgements

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

References

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

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

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

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

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

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

[7] 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 (ACCV'00)*, January 2000.