Human-Computer Interaction - INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 950-953

Real Time Head Gesture Recognition in Affective Interfaces

Rana El Kaliouby and Peter Robinson
University of Cambridge, Cambridge, United Kingdom
{rana.el-kaliouby, peter.robinson}@cl.cam.ac.uk

Abstract: In this paper we present the affective message box, a dialog box that employs a real time head gesture recognition system as its input modality. Head nods and shakes correspond to "Yes/No" options on the dialog box. In addition, a confidence measure is inferred from a number of parameters extracted from gesture's temporal patterns. While in current dialog boxes the input is either a definite yes / no, confidence measures provide applications with additional input from which an appropriate action can be selected. The head gesture recognition system we present here employs feature point displacements to describe basic head motions, which are in turn used with a state-space model to identify the head gesture. Our system detects head nods (95% accuracy) and shakes (91.67% accuracy) in real time (30 frames/sec) without the need for manual pre-processing or prior training.

Keywords: affective computing, head gestures, real time facial expression analysis, feature point tracking

1 Introduction

Facial expressions are often thought of as projections or "read out" of a person's internal mental state (Baron-Cohen et al., 2001). We use facial expressions to express our emotions, but also to provide important communicative cues during social interaction. With the increasing complexity of interfaces and the heightened user expectation of technology, there is a growing need for technologies that recognize affective cues from the user. It is therefore not surprising that more and more researchers are concerned with building systems that are able to model, recognize and respond to the user's affective state (Picard, 1997).

The facial action coding system (Ekman and Friesan, 1978) provides an enumeration of all action units of a face that cause facial movements. There are 44 action units that account for change in facial expressions and twelve that describe changes in head orientation and gaze. A head nod is a series of vertical up (action unit 53) and down (action unit 54) movement of the head used to show agreement or comprehension while listening. A headshake, on the other hand is that motion of rotating the head horizontally from side-to-side (action units 51 and 52) to disagree, or to show misunderstanding of a speaker's words. In an emotional conversation, it is thought to express disbelief, sympathy, or grief (Human Emotions Ltd, 2002).

In this paper, we present the affective message box, which is designed to respond to explicit head gestures from the user. It is built on a real time head unit detector. Head motion such as head-up, head-down, head-turn-left, and head-turn-right are detected and then passed on to a classifier that looks at the temporal relationship between consecutive units and decides if a nod or shake has occurred.

2 Related Work

Head gesture recognition systems deal with three main problems. First, the face region needs to be identified. Tang and Nakatsu (2000) use skin colours to locate a face in a cluttered background. Once the head is localized a feature set thought to represent the gesture is extracted. Tang and Nakatsu (2000) use feature point coordinates detected and tracked over successive frames, while Morimoto et al. (1996) use three image rotation parameters. Kapoor and Picard (2001) and Davis and Vaks (2001) both use the position of the pupils to determine the direction of head movements. Erdem and Sclaroff (2002) use a 3D head tracker to recover head rotation and translation parameters. A classifier then classifies the feature set into one of a possible set of head gestures.

Approaches to classification differ in the number of gestures supported and whether prior training is required. Tang and Nakatsu (2000) use a neural network, trained in advance to classify nods and shakes. Hidden Markov Models (also require prior training) have been used by Morimoto et al. (1996) to classify four gestures (Yes, No, maybe, Hello) and by Kapoor and Picard (2001) to classify nods and shakes. Kawato and Ohya (2000) detect nods and shakes based on pre-defined rules applied to the positions of "between-eyes" in consecutive frames. Davis and Vaks (2001) use two finite state machine models to represent each gesture. The system is tested in a text-editor that employs a dialog-box agent.

Additional issues to be considered in the approach adopted in recognizing head gestures include whether the system operates in real-time and if it's resilient to the user's pose. Prevalent systems range from 13 fps (Kawato and Ohya, 2000) to 30 fps (Kapoor and Picard, 2001), and most require that the user assume a frontal view.

We present a robust head gesture recognition system that is capable of recognizing head nods and shakes without the need for manual pre-processing or prior training. Commodity hardware is used, such as a commercial digital camcorder connected to a standard PC. Our system detects head nods and shakes in real time (30 fps) and doesn't impose a frontal position on the user.

3 Head Gesture Detection

Natural head gestures follow some pattern of temporal regularity (Davis and Vaks, 2001). We utilize those patterns in our system to recognize spontaneous head nods and headshakes, and determine their intensity.

3.1 Head Gesture Parameters

We model head gestures as a set of head motion states and a number of parameters. Head nods are typically characterized by alternating consecutive head-up head-down motions, as illustrated in Figure 1. Headshakes on the other hand are determined by alternating head-turn-left, head-turn-right motions, shown in Figure 2.

Head nods and shakes vary in duration as well as in intensity. Such variations often signify different user intents. For example, a quick and strong nod tends to indicate more agreement than a weaker and slower one. We define a number of parameters that are used to estimate the strength of a head gesture:

- Duration: total span of the gesture
- Peak Displacement: maximum head motion incurred during the gesture.
- Total Energy: the total amount of kinetic energy throughout the gesture.
- Energy rate: the total energy divided by the gesture duration.

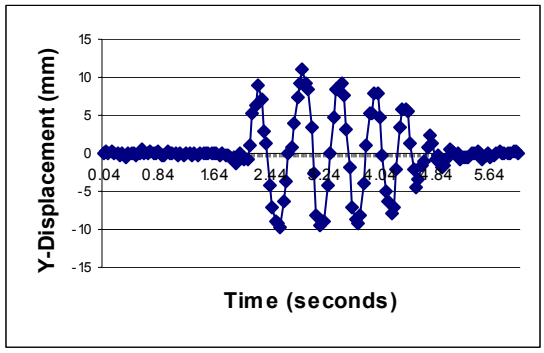

Figure 1: Example head nod determined from the y-displacements of the nose tip point. Nod from 2.12s to 4.84s. Positive displacements indicate an upward motion.

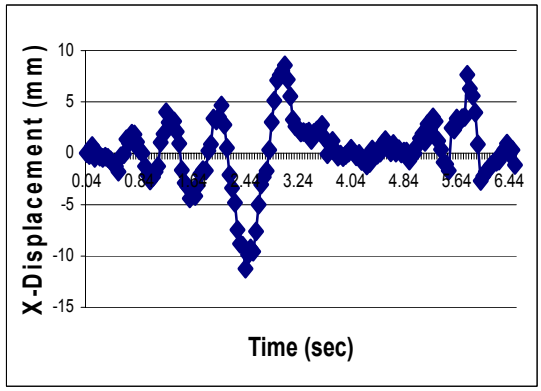

Figure 2: Spontaneous headshake using x-displacements of nose tip point. It starts at 0.24s and ends at 3.88s. Positive displacements indicate a head-turn-right.

3.2 Head Gesture Recognition System

We use a real time face feature tracker to deal with face localization and feature extraction in spontaneous expressions. Our tracker extracts the position of 22 facial features from the video stream.

The y-displacement of the nose tip point is used to distinguish between upward (positive) and downward (negative) motions. The x-displacement differentiates between a turn-left (negative) and turn-right (positive) motion.

The detected head motions are accumulated and are then processed by a nod or shake state machine. The state machine parses through consecutive motions, checking their duration and intensity to decide if they

constitute a valid head gesture. The head nod state machine, shown in Figure 3, consists of four states (start, up, down, finish). A nod can start with an upward or a downward motion. If no adjacent motion is found, the state machine returns to the start state. The head nod state machine looks for consecutive alternating head up, head down motions. A successful nod is returned when the minimum number of alternating motions is detected. In addition, a minimum threshold for purposeful head motion is imposed on the x and y displacements.

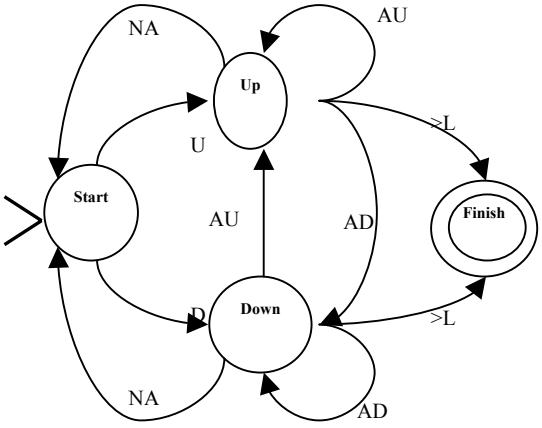

Figure 3: Nod state machine. NA=Non-adjacent motion, U=Upward, D=Downward, AU=Adjacent Upward, AD=Adjacent Downward, L=Length.

4 Experimental Evaluation

We used videos from "Mind Reading", a computer-based interactive guide to emotions (Human Emotions Ltd., 2002) in addition to real time clips. A total of thirty two video clips (recorded at 30 fps) were evaluated:

- 20 nods of varying strengths and durations.
- 12 shakes of different strengths and durations.
- 10 miscellaneous clips (smiling, thinking, and undecided).

The results are summarized in Table 1, and show an overall recognition accuracy of 93.3%. A minimum intensity threshold (to disregard subtle un-purposeful movements), and a maximum duration threshold (to disregard very long head motions in one direction) are used to minimize the number of false nods and shakes.

We found that head motions with speeds higher than that of the video camera's frame rate (30fps) or movements that were outside of the capture window caused the tracker to fail. Our tracker automatically re-locates the face, and restarts the head gesture state machine.

	Correctly Detected	Not detected	False Nod	False Shake	% age
Nod (20)	19	1	0	1	95
Shake (12)	12	1	2	0	91.6
Misc (10)	0	0	2	1	-
	Average recognition accuracy %				**93.3**

Table 1:Summary of evaluation results

5 The Affective Message Box

We use the head gesture recognition system to implement the affective message box. The dialog box responds to head nods for YES, and shakes for NO, instead of keyboard and mouse inputs. This alternate mode of input can be very useful when a hands-free input modality is needed. In addition, using information such as gesture intensity and duration provides applications with a confidence measure, in contrast to a definite Yes or No. This can then be utilized to select an appropriate course of action.

Applications using the affective message box need to be able to determine when to consider a head gesture as a valid reaction. Users might choose to delay the response to a dialog box, or might not be attending to the box. One possible solution involves the use of a "magic" gesture (e.g. looking up to the camera or winking) that would signal the start of a response. In addition, integrating temporal context El Kaliouby et al., 2003) in the inference process can indicate when is an appropriate time to consider a gesture.

Other issues include finding an appropriate head gesture or facial expression for the "cancel" button found on some of the standard dialog boxes. This will be addressed in future work.

6 Conclusion

In this paper we presented the affective message box, a dialog box that employs a real time head gesture recognition system as its input modality. Head nods and shakes correspond to "Yes/No" options on the dialog box. In addition, a confidence measure is extracted from the gesture's temporal patterns. An overall accuracy of 93.33% is achieved. Future work

includes more real time testing, and incorporating other head gestures such as head tilts.

We believe that unobtrusive systems that work on real time video without the need for any preprocessing bring us closer to building affective user interfaces.

References

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001) 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.

Davis, J and Vaks, S (2001). A Perceptual User Interface for Recognizing Head Gesture Acknowledgements, in *Workshop on Perceptive User Interfaces*.

Ekman, P & Friesen, W (1978). Facial Action Coding System (FACS): Manual. Consulting Psychologists Press, Palo Alto, CA, USA.

El Kaliouby, R., Robinson, P & Keates, S (2003) Temporal Context and the Recognition of Facial Expression from Emotion in *Proceedings of the HCII2003 nternational conference on Human Computer Interaction*.

Erdem, U.M and Sclaroff, S (2002). Automatic Detection of Relevant Head Gestures in American Sign Language Communication, in *ICPR'2002: Proceedings of the Sixteenth International Conference on Pattern Recognition*, pp.10460-10463.

Human Emotions Ltd. (2002) Mind Reading: Interactive Guide to Emotion. http://www.human-emotions.com

Kapoor, A and Picard, R.W (2001) A real-time head nod and shake detector, in *Proceedings from the Workshop on Perspective User Interfaces.*

Kawato, S and Ohya, J (2000). Real-time Detection of Nodding and Head-shaking by Directly Detecting and Tracking the "Between-Eyes", in *FG2000: The 4th Int. Conf. on Automatic Face and Gesture Recognition*, pp.40-45.

Morimoto, C., Yacoob, Y, and Davis, L (1996) Recognition of Head Gestures using Hidden Markov Models in nternational Conference on Pattern Recognition, pp. 461-465.

Picard, R. (1997) *Affective Computing*. MIT Press.

Tang, J and Nakatsu, R (2000) A Head Gesture Recognition Algorithm, in ICMI ICMI 2000: Proceedings of Third International Conference on Advances in Multimodal Interfaces, pp 72-80.