MIT Media Laboratory Affective Computing Technical Report No. 544

Appears in: Proceedings from the Workshop on Perceptive User Interfaces, November 2001

A Real-Time Head Nod and Shake Detector

Ashish Kapoor
Affective Computing, MIT Media Lab
20 Ames Street,
Cambridge, MA 02139
+1-617-253-0384
ash@media.mit.edu

Rosalind W. Picard
Affective Computing, MIT Media Lab
20 Ames Street,
Cambridge, MA 02139
+1-617-253-0369
picard@media.mit.edu

ABSTRACT

Head nods and head shakes are non-verbal gestures used often to communicate intent, emotion and to perform conversational functions. We describe a vision-based system that detects head nods and head shakes in real time and can act as a useful and basic interface to a machine. We use an infrared sensitive camera equipped with infrared LEDs to track pupils. The directions of head movements, determined using the position of pupils, are used as observations by a discrete Hidden Markov Model (HMM) based pattern analyzer to detect when a head nod/shake occurs. The system is trained and tested on natural data from ten users gathered in the presence of varied lighting and varied facial expressions. The system as described achieves a real time recognition accuracy of 78.46% on the test dataset.

Keywords

Head Nod, Head Shake, Pupil Tracking, HMM.

1. INTRODUCTION

A very large percentage of our communication is nonverbal, which includes all expressive signs, signals and cues that are used to send and receive messages apart from manual sign language and speech. Nonverbal gestures perform a number of different functions [1]. Head nods and shakes can be used as a gesture to fulfill a semantic function (e.g., nod head instead of saying yes), to communicate emotions (e.g., nodding enthusiastically with approval) and as conversational feedback (e.g., to keep the conversation moving). Table 1 shows some of the semantic functions and emotions associated with head nods and shakes. A system that could detect head nods and head shakes would be an important component in an interface that is expected to interact naturally with people.

Head nod, which is a vertical up-and-down movement of the head rhythmically raised and lowered, is an affirmative cue, widely used throughout the world to show understanding, approval, and agreement [3] [4] [8]. Head shake is rotation of the head horizontally from side-to-side and is nearly a universal sign of disapproval, disbelief, and negation [3] [4] [8]. Although, head nod goes mostly with positive intent/emotions and head shake with negative intent/emotions there might be certain exceptions. For example, head nod might occur with a feeling of rage too. Head nods are also used as a conversational feedback, so a person, even if he does not agree, may nod his head while he is listening.

Table 1. Affective meanings of head nods/shakes

Head Nods	Head Shakes
Approval	Disapproval
Understanding	Disbelief
Agreement	Negation

In this paper we describe a new vision-based system that detects head nods and head shakes in real time. Real time detection of head nods and shakes is difficult, as the head movements during a nod or shake are small, fast and jerky, causing many video-based face trackers to fail. We use an infrared sensitive camera equipped with infrared LEDs to track pupils robustly. A Hidden Markov Model (HMM) [10] based classifier is used to detect the occurrence of head nods and head shakes.

2. RELATED WORK

A lot of work has been done on facial pose estimation as well as on face tracking. 3D facial pose estimation based on facial feature tracking has been suggested [6]. Regions of skin and hair have been used to estimate 3D head pose as well [2]. Rowley et al. [11] [12] have described a neural network based face detector. As the head movements during a nod or shake are small, fast and jerky all these approaches are either unable to track the head in real time during those movements or the resolution they provide is insufficient to detect head nods and shakes.

There is much prior work on detecting and tracking the eyes. Tian et al [13] use a dual state model to recover eye parameters using feature point tracking of the inner eye corners. This system requires that the eye templates be initialized manually in the first frame of the sequence, which prevents it from being automatic. Many eye feature extraction methods are based on deformable templates [14], which are difficult to use for real-time eye tracking and have to be initialized properly to achieve a good performance. Morimoto et al [9] have described a system to detect and track pupils using the red-eye effect. Haro et al [5] have extended this system to highlight the pupils, which in turn are detected and tracked using Kalman filter and probabilistic PCA. Our infrared camera equipped with infrared LEDs, which is used to highlight and track pupils, is an in-house built version of the IBM Blue Eyes camera (http://www.almaden.ibm.com/cs/blueeyes). Our approach to detect and track the pupils is motivated by the methods described in Morimoto et al [9] and Haro et al [5].

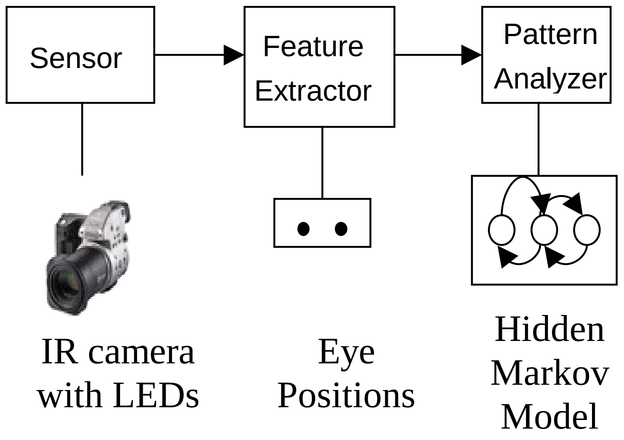


Figure 1. The system architecture

Kawato and Ohya [7] have described a system to detect head nods and head shakes in real time by directly detecting and tracking the "between-eyes" region. The "between-eyes" region is detected and tracked using a "circle frequency filter", which is a discrete Fourier transform of points lying on a circle, together with skin color information and templates. Head nods and head shakes are detected based on pre-defined rules applied to the positions of "between-eyes" in consecutive frames. We describe a simple system that tracks pupils robustly using an algorithm that requires very low processing. However, rather than using a rule-based approach to detect head nods and shakes, we adopt a statistical pattern matching approach trained and tested on natural data. In the next section we describe the infrared camera, the pupil tracking algorithm and the real-time head nod/shake detection based on HMMs. Followed by that we present an experimental evaluation and discuss the recognition results.

3. THE HEAD NOD & SHAKE DETECTOR

Figure 1 shows the overall architecture of the system. An infrared sensitive camera synchronized with infrared LEDs is used as a sensor and produces an image with highlighted pupils. The image obtained from the sensor is processed by the feature extraction module, which detects/tracks the pupil positions and infers the direction in which the head moved. The directions of the head movements in consecutive frames are used as a sequence of observations to train and test the HMMs in the pattern analyzer. The whole system is very efficient and runs in real time at 30fps.

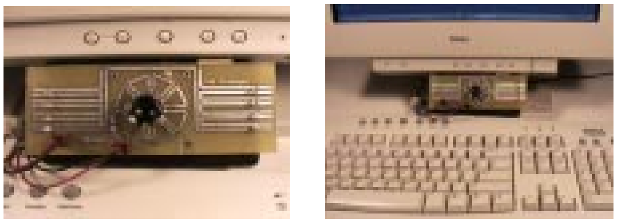


Figure 2. Camera to track pupils, placed under the monitor

3.1 Sensor

The pupil tracking system is shown in Figure 2. The whole unit is placed under the monitor pointing towards the users face. The system has an infrared sensitive camera coupled with two concentric rings of infrared LEDs. One set of LEDs is on the optical axis and produces the red-eye effect. The other set of LEDs, which are off axis, keeps the scene at about the same illumination. The two sets of LEDs are synchronized with the camera and are switched on and off to generate two interlaced images for a single frame. The image where the on-axis LEDs are on has white pupils whereas the image where the off-axis LEDs are on has black pupils. These two images are subtracted to get a difference image, which is used to track the pupils. Figure 3 shows a sample image, the de-interlaced images and the difference image obtained using the system.

3.2 Feature Extraction

The direction of head movement is determined using the positions of the pupils in two consecutive frames. The pupils are detected and tracked using the difference image, which is noisy due to the interlacing and motion artifacts. Also, objects like glasses and earrings can show up as bright spots in the difference image due to their specularity. To remove this noise we first threshold the difference image using an adaptive thresholding algorithm [5]. First, the algorithm computes the histogram and then thresholds the image keeping only 0.1 % of the brightest pixels. All the non-zero pixels in the resulting image are set to 255 (maxval).

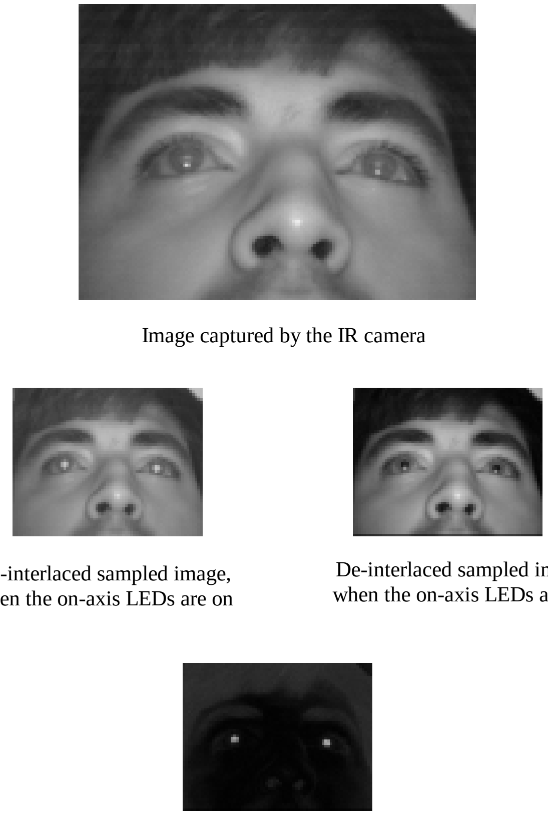


Image captured by the IR camera

De-interlaced sampled image, when the on-axis LEDs are on

De-interlaced sampled image, when the on-axis LEDs are off

The difference image

Figure 3. Pupil tracking using the infrared camera

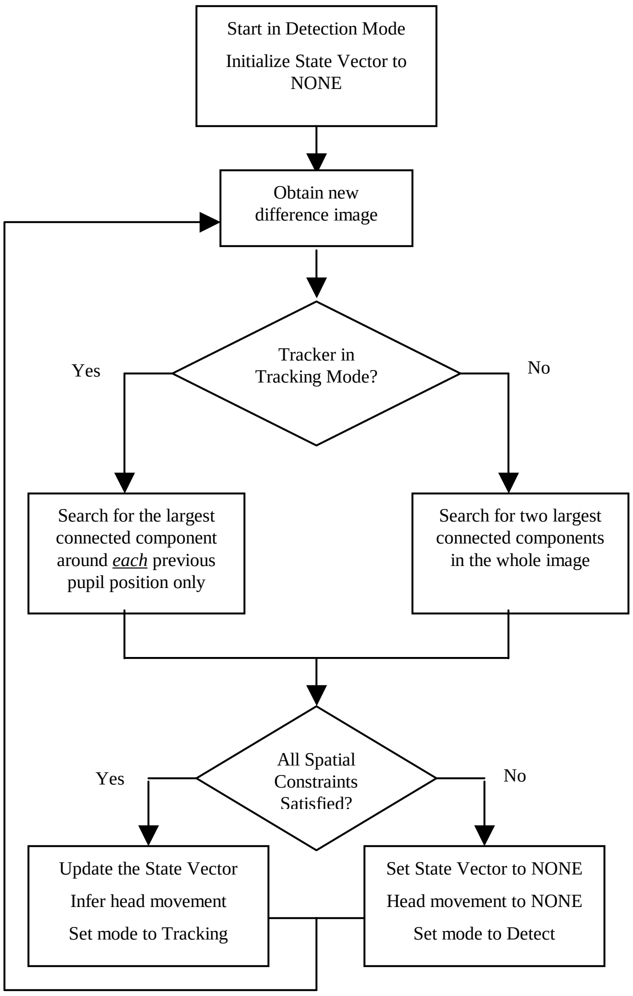

Start in Detection Mode
Initialize State Vector to NONE
Obtain new difference image
Tracker in Tracking Mode?
Yes
No
Search for the largest connected component around *each* previous pupil position only
Search for two largest connected components in the whole image
All Spatial Constraints Satisfied?
Yes
No
Update the State Vector
Infer head movement
Set mode to Tracking
Set State Vector to NONE
Head movement to NONE
Set mode to Detect

Figure 4. The feature extraction module

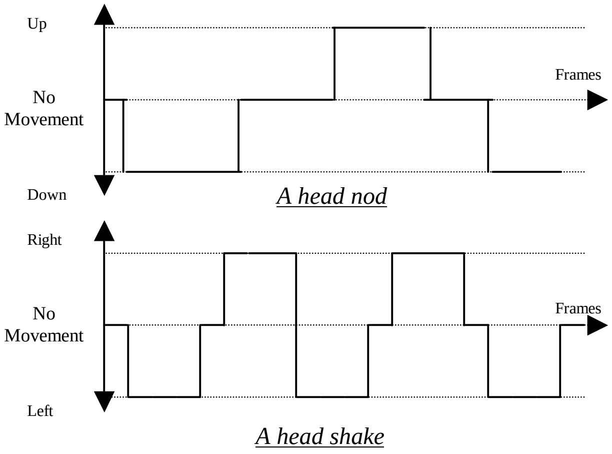
The thresholded image is used to detect and to track the pupil. The pupil tracker is either in a detection mode or a tracking mode. Whenever there is information about the pupils in the previous frame the tracker is in tracking mode and whenever the previous frame has no information about the pupils the tracker switches to the detection mode. The feature extraction module is shown in Figure 4.

During the tracking mode the tracker maintains a state vector, comprised of the spatial information about the pupils. Specifically, the average distance between the pupils during the current tracking phase and their x, y co-ordinates in the previous frames is maintained. To obtain the new positions of pupils a search for the largest connected component is limited to a bounding box centered on previous pupils. The new connected components are accepted as valid pupils when they satisfy a number of spatial constraints. If the area is greater and the displacement of their centers from previous pupil position lies below a certain threshold, the connected components are considered valid. Also if a connected component is found for both the eyes then the distance between these pupils is also compared with the average distance maintained in the state space to rule out false detections. Once the connected components are identified as valid pupil regions, the state vector is updated. Haro et al [5] have used the Kalman filter, which incorporates position and velocity information, to track the candidate regions. Since the motion in the case of a head nod or shake is jerky, we refrain from using a Kalman filter that does not incorporate acceleration and other higher derivatives of position to track the pupils.

The tracker switches to the detection mode whenever there is no information about the pupils. In this mode the tracker simply selects the two largest connected components that have an area greater then a certain threshold. Again, to validate the regions, we apply some spatial constraints. This approach allows us to track the pupils efficiently. Head movements during head nods and head shakes do produce motion artifacts but due to the nature of our algorithm to spatially constrain the search space, it tracks the pupils well. In extreme cases when head movements are too fast, the pupils are lost as motion artifact overpowers the red-eye effect and the pupils are absent from the difference image altogether. For the purpose of detecting head nods and head shakes we found this tracking algorithm to be fairly reliable.

As mentioned earlier we use the direction of head movements as observations for the pattern analyzer. The feature extraction module tracks the pupil positions and based upon that it generates the observations. There are five observation symbols, which correspond to the head moving up, down, left, right, or none. Current pupil positions are compared with pupil positions in the previous frame. If the movement in the x direction is greater than the movement in the y direction then the observation symbol is labeled as left or right head movement depending upon which direction the head moved. Similarly if the movement in the y direction is greater then the movement in the x direction then the label is either up or down, depending upon the direction of the head movement. When the movements in both the x and y directions are below a certain threshold, then the symbol corresponding to none is generated.

3.3 Head Nod & Shake Detection



A head nod

A head shake

Figure 5. Typical sequences of head movements in a head nod and a head shake

Figure 5 shows typical patterns associated with the head movements in a nod and a shake. A head nod is a vertical up-and-down movement of the head rhythmically raised and lowered, whereas a head shake is rotation of the head horizontally from side-to-side. We use a discrete HMM [10] based pattern analyzer to detect when a head nod or a head shake occurs.

Our pattern analyzer consists of two HMMs, one corresponding to head nods and one corresponding to head shakes. Both HMMs have three states and the observation set has five symbols corresponding to the head moving up, down, left, right and none. During the training phase sample head nods and head shakes were processed by the feature extraction module to obtain the sequence of observations, which were used to train the HMMs using the Baum Welch algorithm [10]. In the testing phase, the forward-backward procedure [10] is used to compute the log likelihood for a sequence of N consecutive observations based on the two HMMs. We compare and threshold the log likelihood to label the sequence as a head nod or a head shake. The performance of the system depends upon N, which is the number of observations that constitute a sequence to be tested. If N is small, then slow head nods and shakes might not be detected. When N is large, then the detected head nods and head shakes might linger for some time after they end. Our system uses N=10, which we found sufficient to detect slow as well as subtle head nods/shakes.

4. EXPERIMENTAL EVALUATION

To collect a natural database for head nods and head shakes a Microsoft™ agent was programmed to ask a number of factual questions (see table 2), to which the subjects were asked to answer with a head nod or a head shake. We used this strategy to avoid collecting data with exaggerated head nods and head shakes, which people often made when asked to just nod/shake their head in front of a camera.

Table 2. Ten questions asked by the agent

1. Are the instructions clear?
2. Are you male?
3. Are you female?
4. Are you a student at Media Lab?
5. Are you a student at Boston University?
6. Were you born in Boston?
7. Do you like Boston?
8. Do you like weather here in Boston?
9. A terrible thing just happened in Nepal recently. Did you hear about it?
10. {Agent explains the event} Pretty bad isn't it?

Ten subjects, among whom five were male, five female and two of them wore glasses, were recorded using the infrared camera while they interacted with the agent. We expected to have a total of 100 nods and shakes, but there were instances where the subjects responded to a question with nodding/shaking their head twice. Also, some subjects used head nods as conversational feedback to the agent. A total of 110 samples were collected with 62 head nods and 48 head shakes. Lighting conditions varied due to changes in sunlight coming through a window at different times of day and due to the collection of data from subjects in two different rooms. To further complicate the data, a number of different facial expressions and movements like smiles, and frowns were made by the subjects in addition to the nods and shakes. (Sometimes the agent elicited humor or other responses.) A random 40% of the head nods and 40 % of the head shakes were selected for training (see Table 3). The testing was done on the collected video sequences played using a VCR. This allowed us to test for actual real-time recognition accuracy on the whole set at 30fps. The feature extraction module processed the videos as explained earlier and observations from sequences of ten consecutive frames were used by the pattern analyzer to detect head nods and head shakes.

5. RESULTS AND DISCUSSION

The recognition results are shown in Table 4 and 5. The system was implemented on a Pentium-III 933 MHz Linux machine and a real-time recognition rate of 78.46% was achieved at 30 fps for head nods and head shakes in the test dataset. There were no confusions among head nods and head shakes, as the head movements in a head nod are very different from those in a head shake. Most of the head nods and head shakes that went undetected were the samples taken from the subjects that wore glasses. The specular nature of the glasses made it difficult for the pupil tracker to work well. Interestingly on one of the subjects with glasses, the pupil tracker tracked a bright specular point on the glass frame and hence was able to detect most of the head nods and head shakes. One of the head shakes that went undetected was because the subject closed his eyes while making the gesture.

There were some false positives too. Some head nods were detected when the subject started laughing with the head going up and down rhythmically. Sample demonstration movies can be viewed at http://www.media.mit.edu/~ash/PUI01.

Table 3. Details of training and testing data

	Train	Test
Head Nods	25	37
Head Shakes	20	28

Table 4. Recognition results for the training set

	Recognized		
	Head Nods	**Head Shake**	**Misses**
Head Nods	23	0	2
Head Shakes	0	19	1
	Recognition Rate for Head Nods : 92.0 % **Recognition Rate for Head Shakes : 95.0 %** **Combined Recognition Rate : 93.34 %**		

Table 5. Recognition results for the testing set

	Recognized		
	Head Nods	**Head Shake**	**Misses**
Head Nods	30	0	7
Head Shakes	0	21	7
	Recognition Rate for Head Nods : 81.08 % **Recognition Rate for Head Shakes : 75.0 %** **Combined Recognition Rate : 78.46 %**		

6. CONCLUSION

We have described a system for real time detection of head nods and head shakes using pupil tracking. An infrared camera equipped with infrared LEDs is used to track the pupils. The directions of head movements in consecutive frames, which are inferred from the pupil tracking, are used as observations to train discrete HMMs in the pattern analyzer. Upon seeing a sequence of these observations, the pattern analyzer is able to detect head nods and head shakes in real time. The system was trained and tested on a natural database of head nods and head shakes collected using a Microsoft agent that prompted the subjects to nod/shake their heads by asking a series of questions. Recognition accuracy of 78.46% was achieved for head nods and head shakes on the videos in the test dataset streamed at 30 fps.

Future work includes a system for expression recognition, face and facial feature extraction, pose estimation and a system that integrates all these and acts as an interface that recognizes affect.

7. ACKNOWLEDGMENTS

We thank Dave Koons, Ted Selker, Surj Patel and Yue Hann Chin for their help to build the infrared camera sensor. This research was supported by NSF ROLE grant 0087768.

8. REFERENCES

[1] Cassell, J. Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents, in *Embodied Conversational Agents.* Cambridge, MA, MIT Press 2000.

[2] Chen, Q., Wu, H., Fukumoto, T. and Yachida, M. 3D Head Pose Estimation without Feature Tracking, in Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, 1998.

[3] Darwin, Charles (1872). *The Expression of the Emotions in Man and Animals,* third edition. New York, Oxford University Press, 1998.

[4] Givens D. B. Dictionary of gestures, signs & body language cues. http://members.aol.com/nonverbal2/diction1.htm#The NONVERBAL DICTIONARY

[5] Haro, A., Essa, I., and Flickner, M. Detecting and Tracking Eyes by Using their Physiological Properties, Dynamics and Appearance, in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2000.

[6] Heinzman, J. and Zelinsky, A. 3-D Facial Pose and Gaze Point Estimation using a Robust Real-Time Tracking Paradigm, in Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, 1998.

[7] Kawato, S., Ohya, J. Real-time Detection of Nodding and Head-shaking by Directly Detecting and Tracking the "Between-Eyes", in Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, 2000.

[8] Morris, D. *Bodytalk: The Meaning of Human Gestures*. Crown Publishers, New York 1994.

[9] Morimoto, C., Koons, D., Amir, A., Flickner, M. Pupil Detection and Tracking Using Multiple Light Sources. Technical Report, IBM Almaden Research Center, 1998. http://domino.watson.ibm.com/library/cyberdig.nsf/Home.

[10] Rabiner, L. A tutorial on Hidden Markov Models and selected applications in speech recognition, in Proceedings of IEEE, volume 77, number 2, February 1989, 257-286.

[11] Rowley, H. A., Baluja, S., and Kanade, T. Neural Network-Based Face Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 20, number 1, January 1998, 23-38.

[12] Rowley, H. A., Baluja, S., and Kanade, T. Rotation Invariant Neural Network-Based Face Detection, in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998.

[13] Tian, Y., Kanade, T. and Cohn, J. F. Dual-state Parametric Eye Tracking, in Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, 2000.

[14] Yuille, A., Haallinan, P., Cohen, D., S. Feature Extraction from Faces using Deformable Templates. International Journal of Computer Vision, 1992.