2M4-OS-20a-4

Collection and analysis of multi-party interaction data for boredom recognition

Nataliia Biriukova[*1] Koutaro Funakoshi[*2] Koichi Shinoda[*1]

[*1]Tokyo Institute of Technology [*2]Honda Research Institute

In human-computer interaction systems such as tutoring systems or entertainment robots, it is important to keep users' attention and not to get them bored. For this purpose, first such systems should recognize whether users are bored or not. We plan to develop an automatic boredom recognition system in which several non-verbal cues from users such as gestures and facial expressions are captured and utilized. In this paper we report our database collection for this development. It consists of a set of multi-party conversations including a personal robot, recorded by RGB-D camera and microphones. We annotated 'bored', 'not bored', 'cannot say ', and 'face not visible' categories. We found correlation between physical activities of subjects and their boredom states. The lack of body movements during interaction indicates boredom state.

1. Introduction

The most common human-computer interaction style now is the desktop style, in which the interaction is performed through graphical user interfaces, keyboards, and pointing devices. Although it is very useful when interacting with PCs, it is not enough for emerging applications of computers, such as intelligent tutoring systems or social assistants [1].

With recent technology advance, new kinds of computers for those new applications have been developed. In those applications a system needs to understand users ' affective states, such as emotions, interest level, engagement, and boredom. Humans express their affective state in both verbal and non-verbal cues. Several studies (e.g. [2]) have reported that humans mostly rely on non-verbal cues when judging affective states. Non-verbal cues play important roles in affective state recognition.

Different affective states play different roles and several researches have been devoted to recognition of emotions, interest level, and engagement. On the other hand, automatic boredom recognition importance has not been fully explored. When a person is bored during interaction in any area of life, the goals of interaction might not be fully reached.

In this paper we will first review previous studies and their methods for dataset labeling, then describe our dataset and annotation strategy, and report our results.

2. Previous studies

2.1 Affective states recognition

There has been a number of researches dealing with non-verbal communication cues; to detect user's curiosity in customer service application [3], interest detection in one-to-one interaction [4] and in meetings [5]. There also has been boredom recognition researches based on head positions [6] or on postures [7].

連絡先: Nataliia Biriukova, 080 3019 8755, biriukova.n.aa@m.titech.ac.jp

So far, most of those works has focused on only one modality while simultaneous use of multiple modalities have increased recognition accuracy [8]. Some studies have combined one visual modality such as facial expression with one audio modality (e.g. [9]).

2.2 Dataset labeling methods

Most of the labeling methods in affective computing researches has used annotation by judges and questionnaire. Jacobs [6] used their combination to label boredom states. Participants first labeled how bored they were in each video on a 7-point Likert scale, then two judges put one label per video. The two judges achieved an average of 76.9% agreement after the first annotation. They then went back and re-annotated the events where there was disagreement. This improved the agreement to an average of 96.7%.

In Castellano [10], their dataset was annotated in terms of user engagement with a robot by three annotators. Annotators chose one out of three options and the results from each annotator were then compared. A label was confirmed when it was chosen by two or three of the annotators. In case each of the annotators chose a different label, the segment was labeled as 'cannot say' and was not used in their further study.

Our strategy differs from them. We do not use the questionnaire. Aiming for natural interaction, in each phase of their conversation we focused on long-time interaction scenarios where subjects may not be able to correctly report their boredom state.

3. Database

3.1 Data

Database[*1] [12] consists of 60 recordings, in each of which three users interacting with a robot, recorded by RGB-D camera and microphones. The number of subjects in total is 90. Each recording is 25 minutes long. We used Nao robot [13] and employed Wizard-of-Oz (WoZ) technique in which

*1 this paper's notion of participation is different from the participation annotation described in [12]

Fig. 1: 'Gesture Game' scenario

an operator remotely manipulates a robot, controlling its movement, speech, and gestures. During one session all the users can appear in the scene together, in pairs, or alone. They were instructed to behave naturally, free to leave or join the scene whenever they want.

Each group of three users (further called A, B, and C) participated in two different interaction scenarios. First scenario is 'Quiz Game'. In 'Quiz Game ', the robot imagines a word (e.g. 'apple') and answers yes-no questions of users. Users' goal is to correctly guess the imagined word, asking questions and discussing the robot 's answers with each other. The second scenario is a 'Gesture game' (Fig. 1). It is a game in which the robot tries to teach users a set of gestures in English. For example, the robot touches its nose and says 'Nose', asking users to repeat the same gesture. If user's gesture is correct, the robot gives approving comment.

3.2 Annotation strategy

Annotation is conducted by three judges (further called X, Y, Z), two females (X, Z) and one male (Y). In annotation we used 'bored', 'not bored', 'not sure' and 'face not visible' labels. If a state is observable less than 2 sec, it is not labeled. Followings are the description of the four labels:

A) No face visible - The face of the user is turned from the robot for 90 degrees or more, or the face is blocked by the other user.

B) Bored - The user is not active, reacts slowly, or doesn't react at all to the other participants.

C) Not Bored - The user actively participates in the game, reacts to the robot's questions fast, interacts with the robot or the other users energetically

D) Cannot say - It is extremely hard for the judge to put any of the above categories

Figure 2 shows the decision tree used in the labeling. To answer questions 'Subject participates?' and 'Subject looks interested in participating?', judges used the next rules:

1. Subject participates, when he or she:

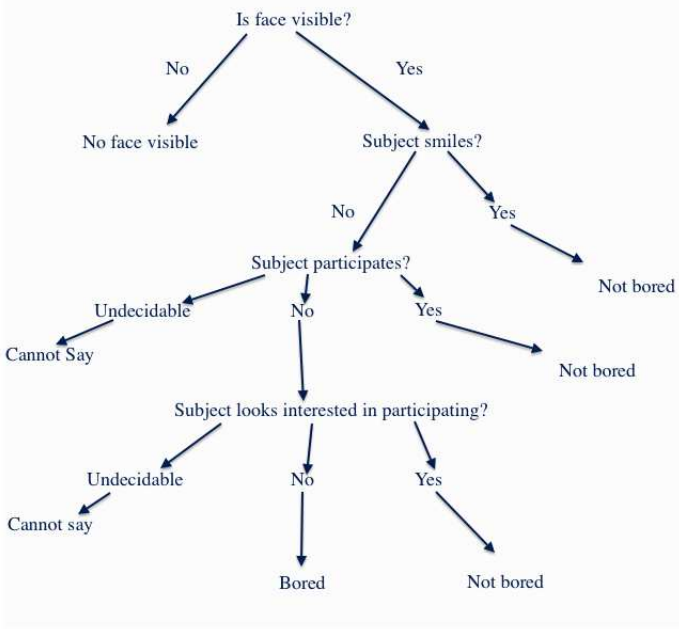

Fig. 2: Annotation decision chart

   (a) Does gestures that the robot asked to do within 3 sec after the robot finished its speech.

   (b) Replies to the questions within 3 sec after the robot finished her speech.

   (c) Raises a hand to reply to the questions within 3 sec after the robot finished its speech.

   (d) Touches or talks to the other subjects.

   (e) Makes excited or happy noises.

   (f) Does not avert his/her gaze from the robot and the other subjects for longer than 7 sec.

2. Subject looks interested in participation, when he/she:

   (a) Looks at the robot or the other participants with smile

   (b) When standing in the back, the subject fixes gaze on the robot or the other participants

   (c) Starts talking to the robot before the robot asks him/her to play

In the cases when judges were not sure about presence of features from the list above and therefore were not able to answer questions in the chart, they annotated ' cannot say ' label.

Some spontaneous gestures are informative for annotators. We listed them in Table 1. When annotators found these gestures, they labeled ' bored ' (gestures from ' Fixing ' group) or ' not bored ' (gestures from the other groups).

Group	Gesture
Fixing	Clothes fixing Hair touching Face touching
Waving	Wave
Win	Win Clap Hands up
Pointing	Pointing Self pointing Next
Playfull	Dancing

Table 1: List of of participants' spontaneous gestures

	A	B	C
Before	7	10	43
After	5	9	13

Table 2: Disagreement rates before and after re-annotation (%)

4. Results

Table 2 shows the disagreement rates for 'Gesture Game' session before and after re-annotation.

Before re-annotation the disagreement rate between judges was high. For example, it was high for C due to his ambiguous behavior. Table 3 shows the example of the amount of time per state, labeled by judge X to three participants, before and after re-annotation. The judge X tended to put 'bored' label more often initially. Also re-annotation reduced the amount of time of 'cannot say' label for all judges. However, it is not clear whether this was due to the better understanding of subjects' reactions or the more biased decisions.

We've found strong correlation between boredom states and the number of spontaneous gestures of subjects. Table 4 shows the amount of gestures in each state for each subject. In 'bored' state subjects tend to be more still and make less gestures. We've also found a correlation between 'bored' state occurrence and the number of participants present in the scene. For cases when only one person interacted with the robot and the person becomes bored, the appearance of the other participants in the scene always causes state change to 'not-bored'. There were no disagreement between judges in all such instances, which makes us to trust the labeling here.

5. Conclusion

An automatic boredom recognition system plays important role in affective state recognition. We collected and analyzed the dataset for such a system, using multiple modalities. We used interactive scenarios for human-robot interaction, recorded the dataset by RGB-D camera and microphones, and labeled them in terms of boredom states. We achieved 80% agreement rate between three judges. We have found the correlation between the physical activity of subjects and their boredom states. The lack of body movements and activeness during the interaction indicates boredom state.

We plan to develop automatic boredom recognition system in future.

	A		B		C	
	Before	After	Before	After	Before	After
Bored	00:57	00:42	01:43	00:50	03:58	03:13
Not Bored	12:47	14:02	11:10	12:23	08:59	10:58
Cannot say	01:09	00:00	01:07	00:21	01:33	00:31
No face visible		00:00		00:05		00:41

Table 3: Time per state before and after re-annotation (min:sec)

	A	B	C
Bored	5	3	11
Not Bored	58	16	60
Cannot say	0	2	5
No face visible	0	0	2

Table 4: Amount of gestures occurred in each state

References

[1] M. Turk, G. Robertson, The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, 2008.

[2] A. Mehrabian, Communication without words, Psychol.Today, vol. 2, no. 4, pp. 53-56, 1968.

[3] P. Qvarfordt, D. Beymer, S.X. Zhai, Realtourist-a study of augmenting human-human and human-computer dialogue with eye-gaze overlay. INTERACT 2005, vol. LNCS 3585, pp. 767-780, 2005.

[4] A. Pentland, A. Madan, Perception of social interest. Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI), 2005.

[5] L. Kennedy, D. Ellis, Pitch-based emphasis detection for characterization of meeting recordings, Proc. ASRU, 2003.

[6] A. Jacobs, B. Fransen, J.M. McCurry, F. Heckel, A. Wagner, J.G. Trafton, A preliminary system for recognizing boredom. Proceedings of the Fourth ACM/IEEE International Conference on Human Robot Interaction, 2009.

[7] S. Mota, R. W. Picard, Automated posture analysis for detecting learner ' s interest level, Computer Vision and Pattern Recognition Workshop, 2003.

[8] E. Hudlicka, To feel or not to feel: The role of affect in human-computer interaction, Int. J. Hum.-Comput. Stud., 59, pp. 1-32, 2003.

[9] M. Pantic et al., Affective Multimodal Human-Computer Interaction, Proc. of ACM Int'l Conf. on Multimedia, pp. 669-676, 2005.

[10] G. Castellano, A. Pereira, I. Leit, A. Paiva, P. McOwan, Detecting user engagement with a robot companion using task and social interaction-based features. Proceedings of the 11th ICMI, 2009.

[11] S.K. D'Mello, P. Chipman, A.C. Graesser, Posture as a predictor of learner 's affective engagement. Proceedings of the 29th Annual Cognitive Science Society, pp. 571-576, 2006.

[12] 石川 真也, 船越 孝太郎, 篠田 浩一, 中野 幹生, 多人数対話ロボットの実現にむけたマルチモーダル対話データの収集と分析著者. JSAI, 2013

[13] http://www.aldebaran-robotics.com/en/