Towards Classification of Engagement in Human Interaction with Talking Robots

Yuyun Huang(✉), Christy Elias, João P. Cabral, Atul Nautiyal, Christian Saam, and Nick Campbell

The School of Computer Science and Statistics,
Trinity College Dublin, Dublin, Ireland
{huangyu,eliasc,cabralj,nautiyaa,saamc,nick}@tcd.ie
http://www.tcd.ie

Abstract. In this paper we describe ongoing work to develop an engagement classifier for human-computer interaction systems. We have successfully classified group and individual engagement in a corpus of a conversation among four people called TableTalk, by using a classifier trained with the Support Vector Machine method and audio-visual features. The goal in this paper is to extend that work for the classification of engagement in videos of interaction between an human and a talking robot. For that purpose we are using a corpus of dialogues between participants and a Lego robot named Herme, which was collected during an exhibition. We describe the techniques to improve the engagement detection by taking into account the differences between the characteristics of the videos between the two datasets. Currently we are also conducting an experiment to manually annotate the Herme videos with engagement labels. These annotations will be used for evaluation and further improvements to engagement detection.

Keywords: Robot interaction · Engagement detection · Voice quality · Visual analysis

1 Introduction

Interaction with non-human interlocutors has become common in scenarios such as talking to an automated banking service over the phone, an enquiry service of a telephone company or even talking to an automated bill payment service. However, in service-oriented scenarios the social involvement in the conversation is not usually important. Our aim in this work is to measure engagement in a social interaction scenario, that is, an human-robot conversation in a public space. There have been previous attempts to model engagement in a similar context. It has been modelled in dynamic environments, where people interact with the system and with each other in a natural manner and they may leave conversations at any time [1].

Both video and audio analysis can be used to model active listening. In [2] engagement is measured in television viewers, by using head pose, five facial

C. Stephanidis (Ed.): HCII 2015 Posters, Part I, CCIS 528, pp. 741–746, 2015.
DOI: 10.1007/978-3-319-21380-4_125

points, head size, and head position. In terms of speech analysis, F0 and other prosodic cues are usually employed for this type of measurement. For example, several prosodic parameters including change in syllabicity, pitch slope and loudness were analysed on non-lexical response tokens in Swedish [3]. The fusion of audio-visual features have also been used in this context, such as for detecting high-interest levels in group meetings [4] and entrainment between members of a group conversation [5].

Recently we have combined visual and speech parameters for the detection of engagement of a group of people talking around a table, from the video recordings and annotations of the TableTalk corpus [6]. This work is going to appear published elsewhere. Interestingly, we have found that voice quality parameters, which are not usually used in this type of studies, obtained good results. In this paper we describe a similar engagement detection method using voice quality and visual parameters to be applied to the corpus of human-robot dialogues.

2 Classification of Engagement in TableTalk Corpus

2.1 TableTalk Corpus

TableTalk data was captured by using a fixed omnidirectional camera. The visual feature extraction process is described in [8]. Authors used the Viola-Jones face [9] detection algorithm and a colour-based tracking method to estimate the face position and size for each video frame. Face motion estimation was performed using a subpixel block matching algorithm. Body and head activity were also measured by calculating the Sum of Absolute Differences (SAD) between the current and previous frames on the body and head regions respectively.

The corpus also contains annotations of engagement ("engaged"/"non-engaged" labels) which were performed by five psychology students [6].

2.2 Engagement Detection

In our previous work of engagement classification, we trained three classifiers using the parameter data sets: audio, visual and audio-visual. After the audio and visual features were extracted they were aligned in time and combined to obtain the speech-visual feature matrix.

The features extracted from the speech signal consisted of the prosodic parameter F0, Mel-frequency cepstral coefficients (MFCCs) and glottal parameters. The glottal parameters were the open quotient (OQ), return quotient (RQ), and speech quotient (SQ). These parameters and F0 were estimated using the method described in [7], while MFCCs were calculated using the SPTK toolkit.

In addition to the visual measurements available in the corpus, we also computed the distance between the positions of the face and the variation in face size between consecutive windowed segments. In total, the number of parameters was six: face movement (distance measure), head forward/backward movement, head horizontal motion, head vertical motion, body activity, and head activity.

We used the engagement annotations and the feature matrix for the classification of engagement with the Support Vector Machine (SVM) method. For testing we used different combinations of audio-visual parameter sets and a 10-fold cross validation approach. The highest accuracy of 76.1 % was obtained for a set of visual parameters combined with the voice quality parameters (MFCCs and glottal parameters).

We also performed a similar experiment for classification of individual engagement, but using the visual parameter set only. The average accuracy rate for the four speakers was 68.7 %. In the future, this work could be extended to incorporate speech features by performing speaker segmenation of the recording and using some method to deal with segments of overlap speech (e.g. applying a source separation technique or simply discarding them). The main limitations for the segmentation are that the automatic speaker separation using signal processing may not be accurate enough and the manual annotation alternative is very time consuming.

3 Detection of Engagement in Human-Robot Dialogues

3.1 Database of Spontaneous Dialogues

A database of human-robot dialogues was collected by using a conversational robot called Herme developed at the Speech Communication Lab of Trinity College Dublin [10]. In that experiment, Herme started conversations with random visitors during an exhibition (HUMAN+ event) that took place at the Science Gallery in Dublin, in 2011 (from April 15 to June 24). In the experiment speakers could move around while interacting with Herme or leave at any point. The "Herme's database" consists of 433 recorded conversations, with clearly included consent form id-number, collected over the three months.

The conversations were recorded from multiple angles using two Sennheiser MKH60 P-48 shotgun microphones mounted at the top of a television screen, which displayed in real-time a top-down view of the interaction. Herme was used together with an auxiliary Lego robot (they looked similar), of which the Herme's webcam was intended to capture the face of the main interlocutor while the webcam on the other robot recorded a more comprehensive scene of the conversations. An i-Sight camera was also used to gather an overall view for the remote operator. Herme was equipped with software for performing face tracking and to move forward/backwards or left/right so that it could keep facing the person.

3.2 Engagement Annotation

We are conducting an experiment for annotation of engagement in the Herme data. The engagement annotation scheme we propose is based on the following four levels of engagement: *high*, *regular*, *low*, and *not-engaged*. In the first case, there is high involvement in the conversion and the person interacts actively (e.g. using body gesture, facial expression and the voice). In this case the person may talk to someone else about the reactions of the robot but they continue to talk with the robot after a few seconds. An example for the second case is when individuals may start talking with each other about a topic unrelated with the interaction with the robot or they may be doing other activities simultaneously (e.g. reading instructions, using mobile phone, etc.). In the third case, the human-robot conversation is considered to be less frequent and participants can be more attracted to the interaction with third-parties or devices than the interaction with the robot. Non-engagement would occur when individuals are not involved in the conversion with the robot at all but they are within its close range.

Five raters will take part in the annotation task and they will be provided with the videos recorded with the robot-mounted camera and the fixed camera beside the robot.

3.3 Audio-Visual Analysis

The speech parameters are the same as those analysed in the TableTalk experiment (described in Sect. 2.2). However, for analysing Herme data we use a voice activity detector because the annotation of non-speech segments is not available (unlike in the previous experiment). Currently we use the VAD of AMR Floating-point Speech Codec [11] reference implementation. This VAD is suitable for real-time applications and we have already used it for demonstrations of Herme where people can interact with the robot.

The visual analysis can be performed on the videos from the static cameras (on the TV and the auxiliary robot) or the camera mounted on top of the moving robot (Herme). We prefer to use the videos from the mounted camera because this is more similar to the type of data that needs to be processed in typical applications of a talking robot with mobility.

In the TableTalk corpus the visual analysis method assumes that changes in the scene are only possible due to the human motion. In contrast, in the Herme data captured from the moving camera changes may occur not only due to body/head movement but also due to the camera motion and other uncontrolled factors of the scene (e.g. the movement of multiple individuals captured by the camera). However, the visual analysis method used for the TableTalk corpus does not take into account the parameter variation caused by the camera motion. We have implemented a modified version of this method that uses the information of the camera motion to analyse the following face movement features: face distance measured between contiguous video segments, head forward/backward and head horizontal/vertical motions.

3.4 Classification of Engagement

In the TableTalk corpus the face detection seems to be accurate, which is expected because the scenario is fixed and the interlocutors are sitting around the table. Sometimes the faces are not detected but this happens just for a small fraction of the video. In the Herme data, face tracking is less reliable due to an higher variability of the visual parameters during short periods of the interaction and sometimes faces are not captured by the camera or just part of the face appears in the video. Also, in this experiment a person may be looking at Herme without speaking for relatively long time, while in TableTalk the silence periods during the dialogue are relatively short. In order to take into account these characteristics of Herme data, we divide the training data into four parts to build different classifiers:

- Segments with voice and face detected are used to train a classifier using audio-visual parameters.
- Segments with voice detected only are used to train a classifier using speech parameters.
- Segments with face detected only are used to train a classifier using visual parameters.
- Segments without voice and face detected are used to train a classifier for idle mode using audio parameters (no human-robot interaction).

The idea of training the last classifier for idle mode is to model the characteristics of the surrounding noise when there is no human-robot interaction. These classifiers can be used in Herme for detection of engagement. For example, Herme can take into account this information to help the decision making and generate smart feedback to the user. For that application the classifier should be selected based on the output of the VAD and face detection components.

4 Summary and Future Work

In this paper we describe ongoing work to measure the engagement of a person with a talking robot. Recently we have studied engagement detection in a corpus of free multiparty conversations among four people sitting around a table (the TableTalk corpus). Based on the findings and developments achieved in that work we propose a method to classify engagement in a database of spontaneous dialogues between people and a robot, Herme. We measure engagement using audio-visual parameters. The speech parameters are related to prosody (F0) and voice quality (mel-cepstrum and glottal parameters), while the visual parameters are related to facial movement. An experiment is being conducted for annotation of engagement in the "Herme dataset" using a new four-level scheme, in order to provide finer descriptors than the engaged/non-engaged annotations of the TableTalk corpus. In this work we also modified the visual analysis method used to extract parameters in TableTalk so that it takes into account the movement of the camera mounted on Herme.

Previously, we obtained an accuracy rate of 76 % for detection of group engagement, based on SVM method. In the Herme corpus, we plan to compare the performance of additional classification algorithms. This classification task is expected to be more difficult than in TableTalk because the first was recorded in a public space with more uncontrolled factors that may affect the audio-visual analysis and the performance of the machine learning algorithms. Also the levels of body movement are much higher in the Herme scenario. A more extensive evaluation of engagement classification using the different modalities (speech, visual, and audio-visual) needs to be carried out because it is more frequent in this scenario that only one of the two modalities is available for detection.

As future work, we plan to investigate additional audio-visual features and integrate the engagement detector into Herme.

Acknowledgments. This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL (www.cngl.ie) at Trinity College Dublin.

References

1. Bohus, D., Horvitz, E.: Learning to predict engagement with a spoken dialog system in open-world settings. In: Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 244–252. USA (2009)
2. Hernandez, J., Liu, Z., Hulten, G., DeBarr, D., Krum, K., Zhang, Z.: Measuring the engagement level of TV viewers. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–7 (2013)
3. Gustafson, J., Neiberg, D.: Prosodic cues to engagement in non-lexical response tokens in Swedish, In: DiSS-LPSS, Citeseer, pp. 63–66 (2010)
4. Gatica-Perez, D., McCowan, I. A., Zhang, D., Bengio, S.: Detecting group interest-level in meetings. In: IEEE ICASSP, pp. 489–492 (2010)
5. Campbell, N.: An audio-visual approach to measuring discourse synchrony in multimodal conversation data. In: Henrichsen, P.J. (ed.) Linguistic Theory and Raw Sound. Samfundslitteratur, Frederiksberg (2010)
6. Bonin, F., Bock, R., Campbell, N.: How do we react to context? annotation of individual and group engagement in a video corpus. In: Workshop on Context Based Affect Recognition, Held in conjunction with SocialCom, pp. 899–903 (2012)
7. Cabral, J. P., Renals, S., Richmond, K., Yamagishi, J.: Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In: Workshop on Speech Synthesis, Germany (2007)
8. Douxchamps, D., Campbell, N.: Robust real time face tracking for the analysis of human behaviour. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 1–10. Springer, Heidelberg (2008)
9. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
10. Vaughan, B., Han, J.G., Gilmartin, E., Campbell, N.: Designing and implementing a platform for collecting multi-modal data of human-robot interaction. Acta Polytech. Hung. **9**(1), 7–17 (2012)
11. Mandatory speech codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Voice Activity Detector (VAD), ETSI Standard TS 126 194 V12.0.0 (2014)
12. Hang, J.G., Gilmartin, E., De Looze, C., Vaughan, B., Campbell, N.: Speech and multimodal resources: the herme database of spontaneous multimodal human-robot dialogues. In: LREC, pp. 1328–1331. Turkey (2012)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newslett. **11**, 10–18 (2009)