A New Multi-modal Dataset for Human Affect Analysis

Haolin Wei[1], David S. Monaghan[1], Noel E. O'Connor[1], and Patricia Scanlon[2,⋆]

[1] Insight Centre for Data Analytics, Dublin City University, Ireland
[2] Bell Labs Ireland, Alcatel Lucent Dublin, Ireland

Abstract. In this paper we present a new multi-modal dataset of spontaneous three way human interactions. Participants were recorded in an unconstrained environment at various locations during a sequence of debates in a video conference, Skype style arrangement. An additional depth modality was introduced, which permitted the capture of 3D information in addition to the video and audio signals. The dataset consists of 16 participants and is subdivided into 6 unique sections. The dataset was manually annotated on a continuously scale across 5 different affective dimensions including arousal, valence, agreement, content and interest. The annotation was performed by three human annotators with the ensemble average calculated for use in the dataset. The corpus enables the analysis of human affect during conversations in a real life scenario. We first briefly reviewed the existing affect dataset and the methodologies related to affect dataset construction, then we detailed how our unique dataset was constructed.

Keywords: Spontaneous affect dataset, Continuous annotation, Multi-modal, Depth, Affect recognition.

1 Introduction

The interpretation of human affect plays an important role in our daily interactions, thus it is crucial for a computer to actually interpret these in order to develop a system that can engage a human in a smooth, natural and emotionally coloured way [1][2][3]. This requires rich sets of labelled [4] and application specific data with in a context that occurs naturally in daily-life [5][11].

There has been a growing interest in collecting multi-modal spontaneous affect datasets during last decade [3][4][12][13], however there is still a lack of emotionally rich social interactions that are captured in a natural real-life setting that is outside a laboratory environment. Such a dataset is required in order to develop affective analysis systems that can be used in daily life.

2D visual signals have been some of the most widely used modalities for affect recognition in the literature. Visual cues such as facial expressions and body gestures have been well studied [9][10][7]. However these 2D visual cues are highly

⋆ This work was conducted whilst the author was with Bell Labs Ireland.

H.S. Park et al. (Eds.): HBU 2014, LNCS 8749, pp. 42–51, 2014.

sensitive to the capture environment such as illumination, occlusions and other changes in facial and body appearance such as glasses, mark-up, facial hair and clothes [6]. Moreover, the single 2D visual analysis is unable to fully capture the out-of-plane changes and structure information. For example certain facial actions such as Jaw Clench could be difficult to detect in a 2D view [6] and same body gesture could give different "appearance" when viewed in different perspectives [8]. In order to tackle these problems, 3D visual signals could be used. With the recent availability of affordable depth sensors (such as the Microsoft Kinect) offering easy access to 3D data of adequate quality, the depth modality has received a lot of attention as it can be used to improve the results and robustness of an affect recognition system. Bio signals such as eye gaze data, electrocardiogram (ECG), electrodermal activity (EDA), respiration amplitude, skin impedance and skin temperature have also been used for affect recognition. Although these modalities can be adapted when visual and audio modalities are not available, E.g. no visible face or speech production [12], they are usually perceived to be invasive and cumbersome and do not lend themselves to be utilised on a large scale basis [14].

In this paper we present one of the first spontaneous and continuous annotated multi-modal dataset focused on human interaction during a debate. The motivation behind building the dataset is driven by the development of a real-time automatic affect recognition system that could provide instant affective feedback to the users.

2 Related Work

As [4] suggested, in general, there are three types of interaction behaviour that have been used to capture human affect i) posed behaviour, where the participant is asked to perform a certain affective state such as happy, sad etc. ii) induced behaviour, where the participant is put in a controlled environment to elicit a certain affective state through various tasks such as watching movies or pictures. iii) spontaneous behaviour, which appear in real-life setting such as debates or other interactions that involve humans and/or machines. Among all three types of interaction scenarios, the posed affect is the easiest to design and capture. However, it have been proven that the affective state raised from a real-life context are more complex compare to the posed ones, as actors tend to exaggerate the affective state they are displaying [14]. Although the induced affective state could provide natural emotional response, it is usually not able to cover the full range and complex affective state as the interaction is restricted to a specific context [3][4]. Finally the spontaneous affect state is the hardest to capture as true affective state are relatively infrequent, short lived, and consist complex context-based changes [14]. Furthermore, by informing a participant that they are being recorded lead to a change in natural behaviours. However, by not informing participants that they are being recorded raises a myriad of ethical issues. In order to ethically capture the spontaneous affective state, various techniques have been developed: i) In [18] the author use a series of activities

such as, listening to a joke or experiencing harsh insults from the experimenter to try to elicit target emotional state. ii) In [4] the authors use the survival task techniques where group discussion is promoted by asking participants to reach a consensus on how to survive in a disaster scenario.

Various multi-modal datasets that consist of spontaneous and socially enriched human affective states have been created in the last decade. The SEMAINE [3] dataset was one of the first corpus focused on machine-human interaction using nonverbal expressions. It was released in 2007 and is one of the most widely used affect dataset for benchmarking human affect recognition systems [3][16][17]. It features continuous annotated audiovisual recordings of emotionally coloured conversations, elicited through a Sensitive Artificial Listener(SAL). There are four SALs where each SAL is designed to drive the user towards a specific affective state (angry, happy, gloomy and sensible) using predefined subscripts. The annotations include five core dimensions (valence, activation, power, anticipation/expectation) and optional descriptors such as basic emotions and epistemic states.

The Cam3D [13] corpus is a 3D multi-modal corpus which consists of elicited complex mental states. The dataset includes 12 mental states (thinking, concentrating, unsure, confused, triumphant, frustrated, angry, bored, neutral and surprised) which was captured using two High-definition (HD) cameras and two Kinect sensors. The community crowd-sourcing was used to annotate the data.

The MAHNOB-HCI [12] is a multi-modal dataset including synchronized recordings of video, audio, eye gaze data and physiological signals. The emotions are elicited by watching various video clips with different emotional keywords. The data was annotated with emotional tag (neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise, and fear) and a 9-Likert scales on the arousal, valence, dominance and predictability dimensions using self-assessment.

The RECOLA corpus [4] contains spontaneous collaborative and affective interactions in French. The dataset was recorded in dyads during a video conference while completing a task requiring collaboration. The recordings include video, audio, ECG and EDA. The data was continuously annotated on valence and arousal dimensions. Additionally, a 7-Likert scale was used to describe the social behaviours on the five following dimensions: agreement, dominance, engagement, performance and rapport.

The EAGER Spontaneous 4D-Facial Expression Corpus [18] is the latest high-resolution spontaneous 3D dynamic facial expression dataset. The target emotional expressions include happiness, sadness, surprise, embarrassment, fear, physical pain, anger and disgust. The dataset consists of high resolution texture and depth, however it have not been made public available yet.

Although various corpus have been created, to our best knowledge, there does not exist any corpus that includes recording of spontaneous behaviours with both audiovisual and depth data that is also annotated continuously in multi-dimensional affective space. Another novelty of our multi-modal dataset is the chosen scenario: three way debate, this specific scenario was designed to replicate the commonly used Skype or Google hang-out type multi-video

conversation. The scenario will not only allow the researcher to study the spontaneous affective state in relation to different modalities, but also enable the study of affect response between different participants. In addition, instead of capturing the dataset in a controlled laboratory environment, the capture is performed at various locations with different lighting conditions. A comparison of the recent publicly available datasets with our dataset is shown in Table 1.

Table 1. Overview of Human Affective Datasets. Types: P:posed, I:induced, S:spontaneous. Modalities: V:video, A:audio, D:depth, E:EEG, G:gaze, ED:EDA.

Dataset	Types	Subjects	Duration	Modalities	Annotation
SEMAINE (2008)[3]	I	20	04:11	V/A	Continuous
EAGER (2013) [18]	S	41	05:28	V/D	Discrete
Cam3D (2011) [13]	I	16	06:00	V/A/D	Discrete
MAHNOB-HCI (2011) [12]	I	27	06:00	V/A/E/G	Discrete
RECOLA (2013) [4]	S	46	06:30	V/A/E/ED	Continuous
DCU Affect Dataset	**S**	16	**05:30**	**V/A/D**	**Continuous**

3 Dataset Construction

3.1 Affective States Elicitation

In order to capture the spontaneous affective state, the debate scenario contained the following attributes: i) compared with other scenarios, the debate occurs naturally in everyday life, such as in a meeting, when watching a football match or movies. Participants are moved by real motivations leading to highly spontaneous affective state. ii) debate scenarios convey rich affective state and social behaviours such as conflicts, dominance, agreement/disagreement and interest/non-interest [23]. The following topics were selected for the debate:

1. How Ireland performed in the Six Nations Rugby match.
2. Should Ireland reduce the minimal wage?
3. Will Irish economy take off in the future?
4. Do humans have free will?
5. Do humans have a moral obligation to be vegetarian?

The first topic was used in two sections. The first section consists of three sports fans, allowing the capture of strong interest. The second section includes two sports fans and one non-sports fans, this will ensure the capture of rich interest and non-interest. The rest of the topics were used to enable the capture of agreement/disagreement, dominance, positive and negative valence.

3.2 Participants and Environment

16 participate from Dublin City University and Bell Labs Ireland were recruited for the dataset capture. The 16 participants consisted of 3 females and 13 males with an age group ranging from 20 to 50 years old.

Six offices with various background and illuminations were used during the dataset capture. In order to capture both facial expressions and upper-body gestures, each participant was arranged to sit 1 meter away from the screen.

3.3 Procedure

Each debate section consisted of 3 participants in which the 3 participants were first introduced to each other, then they were separated in three different offices and received an introduction on the experiment and an debate topic. Similar to [13], a wizard-of-oz method was used. Participants were told at the beginning of the experiment that their video and audio will be recorded for face and voice recognition. Without knowing the real objective of the experiment will avoid having participant exaggerate or mask their true affect state [13]. Each section will be ended either a time limit is reached (60 minutes) or the debate comes to a natural conclusion.

3.4 Multi-modal Data Capture Equipment Used

Due to the low quality of the Kinect RGB camera, a High-definition (HD) webcam (Logitech C910) was used and placed on top of the Kinect to collect the visual signals at each office. The microphones in the HD webcam were used to capture the audio signals and the Kinect was used to capture the depth information (As Shown in Figure 1).

Two computers were used in each office, one computer is used by the participant to communicate through each other using Google hang-out, while the other computer is used to capture the multi-modal data. The HD webcam provided 1280 x 720 px resolution colour images at 30 frames per second. The Kinect sensor consists of a normal RGB camera and an infrared camera. The RGB camera is able to provide 640 x 480 px color image and the infrared camera is used to capture the structured light and calculate 640x480 px 11-bit disparity map. A headphone was used by each participant to prevent the microphones capturing other participant's voices. In order to reduce the load on the hard drive, only the depth streams from the Kinect were recorded. The audio was recorded using the Microphone on the HD webcam at 16 bit 96kHz. Camera calibration was performed between the HD webcam and Kinect infrared camera in order to map the depth information to the RGB image. Because the video and depth signals are captured from different sensors subsequent manual synchronisation was required. The video stream was compressed using MJPEG. The depth stream was saved in ONI format which was developed and used by OpenNI framework, this allows the use of Natural Interaction for The End user (NITE) library to detect and track upper-body skeleton joints. The audio stream was saved as raw audio (PCM) format. Sample screenshot from the dataset is shown in Figure 2.

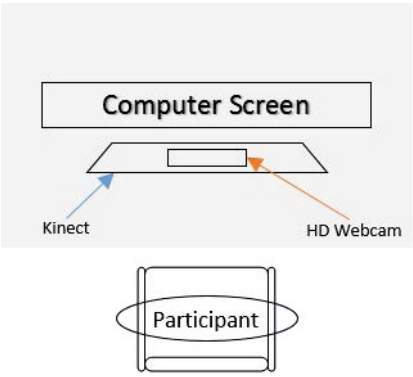

Fig. 1. Capture Environment Layout

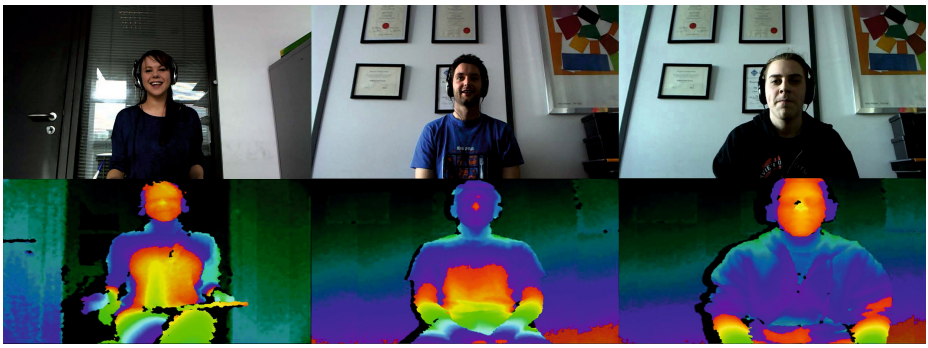

Fig. 2. Sample screenshot from the dataset

4 Segmentation and Annotations

4.1 Segmentation

Due to each debate section usually last from 40 to 60 minutes long, the videos were segmented into 5 to 10 minute clips for easier annotation. We decide to choose the middle part from each section for annotation as the beginning usually consists of warm up chat while at the end of a section people might end up with discuss other topics. This results 34 video clips consists of approximate 5 hours and 30 minutes data.

4.2 Annotation Tool

Currently numerous tools have been developed with different features to annotated different type of datasets. **E**uropean distributed corpora project **L**inguistic

ANnotator (ELAN) [19] is an annotation tool that allow user to create, edit, visualize and search annotations for video and audio data. Another widely used video annotation tool is **AN**notation of **VI**deo and **L**anguage (ANVIL) which was introduced in 2001 [20]. ANVIL is designed to annotate audiovisual material containing multi-modal dialogue. The FEELtrace is an annotation tool developed to enable the raters track the affect state via vocal and visual cues over continuous traces in the dimensional space [21]. FEELtrace allow raters watch the audio-visual recording and rate the perceived emotion sate by moving the mouse pointer within the 2-dimensional of valence-arousal space. The value of the affect state have been confined to [-1, 1] where -1 represent very negative (valence) or very passive (arousal) and 1 represent very positive (valence) or very active (arousal). More recently, the General trace (Gtrace) have been introduced to replace the FEELtrace with the ability to let people use their own dimensions and scales [22]. Due to the simple interface and continuous annotation support, Gtrace was chosen to annotate the data. Figure 3 shows a screenshot from the Gtrace annotation tool used.

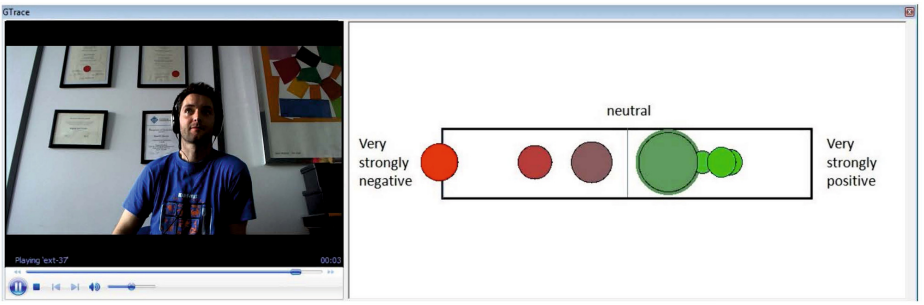

Fig. 3. Screenshot of Gtrace

4.3 Annotation Guidelines

Three independent annotators were hired, before the annotation task, each annotator was briefly introduced to the annotation task. Then they are required to complete a list of training tasks to test their affect recognition skill and to get familiarise with the use of Gtrace. The first task involves the identification of emotions expressions expressed on the face. The second task requires participants describing the emotional state showed in a video clip. The third task involves mapping a list of 24 emotional keywords to a valence-arousal 2-dimensional space. Task 4 involves annotating a list of sample videos from SEMAINE dataset[3] use Gtrace.

Annotation was based on context-free observer judgment, each video clip was continuous annotated in 5 dimensions: arousal, valence, agreement, interest and content. To help the annotator better follow the conversation, the audio from each participant was mixed together.

5 Statistical Analysis

The annotations were first post-processed to remove duplicated annotations and then cropped to temporally align with the video sequences. For comparison purposes, the annotation data was binned with a frame rate fixed to match the video frame rate following the approach used in [4], which is a 33ms duration bin in our case. The percentage of positive frames, mean correlation coefficient and the Cronbach's alpha were computed for each dimension. The correlation coefficient measure the linear dependence between two variables, giving a value between -1 to +1, where 1 indicates total positive correlated, 0 indicates no correlation, and -1 indicates total negative correlated. The Cronbach's αwas used to estimate the internal consistency between annotations where $\alpha>0.7$ is considered as an acceptable internal consistency and $\alpha>0.8$ indicate good consistency. Due to the nature of the debate scenario, the raw data shows higher percentage of positive arousal (compare to the RECOLA corpus) and interest frames. The percentage of positive valence frames is similar to RECOLA corpus with lower internal consistency (See Table 2). The annotation also shows the capture of agreement and disagreement as well as positive and negative content. When annotated the data, the raters show much higher agreement on arousal and interest dimensions compare to valence, agreement and content dimensions (see Table 3).

Table 2. Compression of the statistics of the affective behaviours between RECOLA and our datset

Statics Properties	Arousal (RECOLA)	Arousal (Ours)	Valence (RECOLA)	Valence (Ours)
% Pos Frame	52.1	97.3	75.5	73.3
Mean Corr.	0.435	0.76	0.407	0.47
Mean α	0.80	0.89	0.74	0.66

Table 3. The statistics of the other three dimensions

Statics Properties	Agreement	Content	Interest
% Pos Frame	79.6	74.8	94.6
Mean Corr.	0.46	0.39	0.66
Mean α	0.63	0.60	0.83

6 Conclusions

A new 3D multi-modal spontaneous affect dataset has been introduced. 16 participants were recorded during a sequence of debates in a video conference, Skype style arrangement. Recording include video signals, audio signals and depth signals. Over five hours data have been manually annotated in 5 dimensions including arousal, valence, agreement, content and interest. The analysis of the annotations shows a good inter-agreement on arousal and interest dimensions

and acceptable one for valence dimensions. The overarching goal behind the creation of this dataset is to provide a new rich annotated source of data that can be utilised by the research community for work in automatic human affect analysis. The dataset will be made public available for research purposes.

Acknowledgements. We would like to thank Patricia Scanlon and Philip Kelly for their supports during the dataset capture. This work is co-funded by Bell Labs Ireland and the Irish Research Council under the Enterprise Partnership scheme. The research that lead to this paper was also supported in part by the European Commission under the Contract FP7-ICT-287723 REVERIE.

References

1. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18(1), 32–80 (2001)
2. Cowie, R., Schröder, M.: Piecing together the emotion jigsaw. In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 305–317. Springer, Heidelberg (2005)
3. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on Affective Computing 3(1), 5–17 (2012)
4. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8 (April 2013)
5. Cowie, R., Douglas-Cowie, E., Martin, J.-C., Devillers, L.: The essential role of human databases for learning in and validation of affectively competent agents. OUP, Oxford (2010)
6. Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L.: Static and dynamic 3d facial expression recognition: A comprehensive survey. Image Vision Comput. 30(10), 683–697 (2012)
7. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. IEEE Transactions on Affective Computing 4(1), 15–33 (2013)
8. Aggarwal, J.K., Xia, L.: Human activity recognition from 3d data: A review. Pattern Recognition Letters (2014)
9. Pantic, M., Bartlett, M.S.: Machine analysis of facial expressions. In: Delac, K., Grgic, M. (eds.) Face Recognition, pp. 377–416. I-Tech Education and Publishing, Vienna (2007)
10. Gunes, H., Pantic, M.: Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 371–377. Springer, Heidelberg (2010)
11. Afzal, S., Robinson, P.: Natural affect data; collection and annotation in a learning context. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, pp. 1–7 (September 2009)

12. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing 3(1), 42–55 (2012)
13. Mahmoud, M., Baltrušaitis, T., Robinson, P., Riek, L.D.: 3D corpus of Spontaneous Complex Mental States. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 205–214. Springer, Heidelberg (2011)
14. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image Vision Comput. 31(2), 120–136 (2013)
15. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: 8th IEEE International Conference on Automatic Face Gesture Recognition, FG 2008, pp. 1–6 (September 2008)
16. Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: AVEC 2011–the first international audio/visual emotion challenge. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 415–424. Springer, Heidelberg (2011)
17. Schuller, B., Valstar, M., Cowie, R., Pantic, M.: AVEC 2012: The continuous audio/visual emotion challenge - an introduction. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI 2012, pp. 361–362. ACM, New York (2012)
18. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P.: A high resolution spontaneous 3d dynamic facial expression database. In: Proceedings of 10th IEEE International (2013)
19. Brugman, H., Russel, A.: Annotating multi-media/multi-modal resources with elan. In: LREC (2004)
20. Kipp, M.: Anvil - a generic annotation tool for multimodal dialogue (2001)
21. Schröder, M., Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M.: 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time. In: Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast, pp. 19–24. Textflow (2000)
22. Cowie, R., Sawey, M.: GTrace-General Trace program from Queen's, Belfast (2011), `https://sites.google.com/site/roddycowie/work-resources` (Online; accessed April 29, 2014)
23. Vinciarelli, A., Dielmann, A., Favre, S., Salamin, H.: Canal9: A database of political debates for analysis of social interactions. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, pp. 1–4 (September 2009)