NAA

A Multimodal Database of Negative Affect and Aggression

Lefter, Iulia; Jonker, Catholijn M.; Klein Tuente, Stephanie; Veling, Wim; Bogaerts, Stefan

DOI
10.1109/ACII.2017.8273574

Publication date
2017

Document Version
Final published version

Published in
Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII2017)

NAA: A Multimodal Database of Negative Affect and Aggression

Iulia Lefter*†, Catholijn M. Jonker†, Stephanie Klein Tuente‡, Wim Veling‡ and Stefan Bogaerts§ ¶
*Systems Engineering, Faculty of Technology, Policy and Management , Delft University of Technology
†Interactive Intelligence, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology
‡Department of Psychiatry, University Medical Center Groningen
§Department of Developmental Psychology, University of Tilburg
¶FPC de Kijvelanden

***Abstract*—We present the collection and annotation of a multimodal database with negative human-human interactions. The work is part of supporting behavior recognition in the context of a virtual reality aggression prevention training system. The data consist of dyadic interactions between professional aggression training actors (actors) and naive participants (students). In addition to audio and video, we have recorded motion capture data with kinect, head tracking, and physiological data: heart rate (ECG), galvanic skin response (GSR) and electromyography (EMG) of biceps, triceps and trapezius muscles. Aggression levels, fear, valence, arousal and dominance have been rated separately for actors and students. We observe higher inter-rater agreement for rating the actors than for rating the students, consistently for each annotated dimension, and a higher inter-rater agreement for speaking behavior than for listening behavior. The data can be used among others for research on affect recognition, multimodal fusion and the relation between different bodily manifestation.**

1. Introduction

Forensic psychiatric inpatients are frequently victims of aggressive behavior by fellow patients, as well as perpetrators of aggression. In many cases they are not aware of their own aggression, and when provoked they are not able to de-escalate a conflict [17]. Standard aggression regulation training follows a cognitive behavioral therapy approach. Effects in forensic settings are small, because possibilities for controlled exposure to real-life provocation and practicing new behavior are limited. An interesting alternative is the development of a Virtual Reality Aggression Prevention Training (VRAPT) which is the context of our work. VRAPT offers an interactive three-dimensional virtual world in which social situations and interactions can be experienced and practiced. The patients have to interact with avatars, and in order to make the experience real and increase presence, the avatar has to respond appropriately.

A key component of VRAPT is behavior recognition: the system detects the emotional state and aggressive tendency of the patient. The focus of this paper is the development of the NAA multimodal database, whose primarily purpose is to serve as training material for the behavior recognition component. The dataset consists of human-human interactions showing negative affect and different levels of aggression, excluding physical violence.

We have considered the datasets made available by the research community. Our main requirements were: having data as similar as possible to the final application, with realistic behavior, considering a large range of sensors, having emotionally colored and aggressive content, and scenarios easy to imagine and feasible for aggression de-escalation training. Another important aspect is having different levels of (upcoming) aggression, for which specific reactions of the avatars can be modeled and applied during therapy. Even though we learn a lot from the available datasets, they don't match our criteria fully.

Aggression and negative affect determine changes in facial expressions, body language, speech, as well as physiological changes [9]. Often there is a distinction between reactive and proactive aggression. Reactive aggression is an impulsive and uncontrolled outburst of anger in reaction to (perceived) provocation. On the contrary is proactive aggression, a more planned and well thought form of aggression. The perpetrator has something to gain, for example power or money [7]. Therapists in forensic clinics confirmed that the behavior of inpatients is diverse, some showing overt behavior known as reactive aggression, and some being more instrumental and presenting less visible signs of upcoming aggression. The modalities recorded to account for overt behavior are: audio, video, motion capture (using Microsoft Kinect) and head tracking. To monitor the physiological changes experienced by the subjects we also recorded electrocardiogram (ECG), galvanic skin response (GSR) and muscle activity of biceps, triceps, trapezius (EMG).

Databases have an essential role in training models for recognition systems. For achieving good performance in practice, it is important that the context of the database is similar to the context in which the system will be used [5], and also that the recording conditions are similar [20]. Ideally, data of forensic psychiatric inpatients would be collected in real life settings. Since this is impossible due to ethical concerns as well as practical issues related to real life exposure of the inpatients, we recorded interactions

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

between professional aggression training actors that work with forensic institutions and naive participants. Together with clinicians working in the field we have designed a set of easy to imagine scenarios that can be part of VRAPT. Given these scenarios we expected a range of negative emotions such as impatience, frustration, anger and even aggression.

The remainder of this paper is organized as follows. Section 2 gives an account of related work with respect to other databases. The design considerations of the NAA database including scenarios, recording protocol, and technical characteristics are introduced in Section 3. Section 4 reports on the annotated dimensions and annotation outcome. The paper continues with a discussion on the outcome of recording and annotating the dataset in Section 5 and ends with summary and conclusion in Section 7.

2. Related work

Ongoing efforts in the last decade focus on creating multimodal corpora consisting of affective and social interactions. Such datasets support a broad range of research, such as understanding the relationship between behavioral cues and their communicative function, the development of recognition systems that automatically assess such behavioral signals and their integration in computer-aided systems serving different applications. One of the main points of debate is the realism of the datasets [12], [8], for example with respect to emotional content. In that sense, datasets originally started with posed emotions [2] and evolved by finding methods of eliciting spontaneous content, for example by using a Wizzard of Oz setup for children interacting with the Aibo robot [1] or during interactions with avatars with different personalties in the SEMAINE database [22].

There is an increasing interest for applications of affective and social signal processing in the medical domain. As such, challenges have been organized by the speech community for depression recognition [29] and autism [27]. The work in [6] presents the SimSensei Kiosk, where a virtual human providing health care decision support is developed. The system identifies verbal and non-verbal cues of mental illness and is able to carry out a discussion with the patient autonomously or with limited human support.

The Distress Analysis Interview Corpus (DAIC) [10] is a collection of semi-structured clinical interviews, part human-human and part human-avatar, that was used in creation of SimSensei. The participants are US veterans as well as general public, part of them suffering from depression, PTSD or anxiety. The recordings contain audio-visual and motion-capture data (Microsoft Kinect), as well as physiological data: ECG, GSR and respiration. Even though this corpus contains sufficient sensors, visible signs of distress, and has the advantage of interviewing participants with genuine clinical conditions, it is not suitable for our research due to the lack of aggressive content.

Both audio-visual and physiological sensors are also recorded as part of the RECOLA dataset [25]. It contains spontaneous interactions during a collaborative survival task, where the mood of the participants was or was not manipulated in advance by asking them to watch an emotion eliciting movie. Again, the lack of aggressive content restricts us from using this dataset.

A number of datasets with surveillance applications are presented in [4]. However, they focus on human actions and the recordings are mostly video only. Recordings of aggression in the train and train station domain are used in [21], while the dataset introduced in [18] contains negative interactions in the context of a service desk. The last two corpora are both audio-visual and are based on actor improvisations. These datasets are suitable for us in terms of aggressive content. However, they do not offer the entire sensor range we are interested in.

The IEMOCAP [3] corpus consists of scripted and improvised interaction between actors, and have in addition motion-capture data. Another approach of dataset construction is using TV material. The Vera am Mittag dataset [11] contains recordings from a reality show. The Canal 9 corpus [31] focuses on political debates. While these datasets are interesting from the content point of view, they do not provide the whole sensor range we are interested in.

To conclude, multimodal datasets are in continuous development. Compared to early work, they focus more on realism and spontaneity, include the use of different sensors and focus on general or specific applications. In this context, we present yet another multimodal dataset focusing on specific applications. It can be used for studying overt behavior during negative interactions, and also offers a range of physiological signals of people in unconfortable situations. Having both aggression training actors and naive participants recorded facilitates investigating the differences between acted and spontaneous reactions. The wide range of sensors supports research on multimodal fusion and the relation between different bodily manifestations.

3. Database design

3.1. Recording protocol

The recordings consist of dyadic interactions (in Dutch) between professional aggression training actors and naive participants. To ensure the manipulation of aggression levels during these interactions, 4 aggression training actors from forensic clinics (3 male, 1 female) were hired. They are specialized in showing and eliciting aggression. Their aim is to challenge the trainees to deal with aggression, to give them feedback (e.g. what message their body language was conveying), and also to teach them methods to best behave when confronted with aggression. This is most often done via role-playing. Since we were not allowed to employ forensic psychiatric inpatients, 12 naive participants (3 male, 9 female) were hired to interact with the actors. The actors act as confederates, since their role is not known before hand by the students. For ease of communication, we will refer to the aggression training actors as“actors” and to the naive participants as “students”.

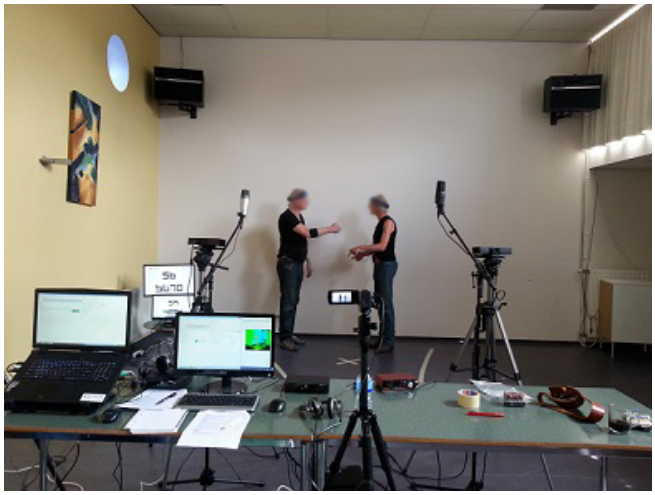

Figure 1. Recording setup.

Both students and actors received separate instructions about the scenarios in advance, and the instructions for each of them specified a goal to be achieved. There were 3 different scenarios in total. The interactions took place according to the following scheme. Each actor played each scenario 3 times, every time with a different aggression level evolution and every time with a different student. First they were asked to start neutral and to escalate, the second time to start highly aggressive and de-escalate, and the third time to keep a high level of aggression during the whole interaction. Each student participated in each of the three scenarios once, every time with a different actor, and every time with a different aggression level variation. Overall, the scheme leads to 36 recorded interactions. The actors were signed when 90 seconds had passed, and the interactions were stopped after 3 minutes if they did not end before that. After the role-plays ended and the participants cooled down, physiological data was recorded again to obtain a personal base-line. Finally, all participants were debriefed and filled in a questionnaire.

3.2. Scenarios

The scenarios had to fulfill a set of requirements. First of all, they had to be easy to imagine and possible to encounter in daily life. Also, they had to be similar to the scenarios that will be used in the VR aggression prevention training of the forensic psychiatric inpatients. The descriptions of the actors' and the students' scenarios consist of a context they are in (involving a degree of urgency), a role (e.g. you are a bus driver), and a goal (e.g. do not let the passenger travel without paying). It is important to note that they were assigned roles only and no actual scripts. This ensures that the interaction builds up naturally in reaction to each other's behavior. A similar technique was used in [18], [21] and [3], with the difference that in these previous works all participants were actors. Table 1 provides an overview of the scenarios' content. In the first two scenarios the student has to forbid the actor something, in the third scenario the student has to make sure the actor does him/her a favor.

3.3. Technical details

Sound was recorded with 4 microphones at a sample rate of 44.1 kHz: 2 omni directional microphones at a distance (Samson C03U), and 2 wireless collar microphones (AKG CG 55). For motion tracking 2 Microsoft Kinect for Windows v2 sensors were used, mounted at an angle such that each of them would capture one participant as close as possible to frontal position, and capture the other participant as well when they are close. This was to diminish the effects of self occlusion when recording participants from the side, and also because in the final setup of the VR aggression prevention training there will be only one participant. The interaction of the two participants is visible, which can be practical for other applications. Frame rate for skeleton and lower resolution video was 25fps.

Each participant was wearing a head band with a Chrum tracker (based on CH-Robotics UM-6 IMU) attached at the back, which recorded Euler angles and quaternions. In the final system the patients will be wearing an HMD with a similar sensor, and this supports the analysis of head motion during interactions. In addition, subtle head movements are not well visible in the skeleton extracted from Kinect, so this allows for more exact head motion tracking.

Physiological data was recorded using MindMedia NeXus sensors: one NeXus 4 and one NeXus 10 were available. Both NeXus types support ECG, GSR and EMG but NeXus 4 has less ports. We used the NeXus 10 for the students because we expected them to behave more spontaneously and naturalistically, and NeXus 4 for the actors. ECG (hear rate and R-R interval) and GSR (index and ring finger) we recorded for students and actors. Muscle activity of the biceps, triceps and trapezius (EMG) was recorded for students and EMG of the biceps was recorded for the actors. The choice of muscles to monitor was based on previous research [30] of how these muscles are activated during emotional arousal. Another motive was that the same muscles should be used during therapy, and placing electrodes on some other body parts might be troublesome for forensic psychiatric inpatients. Physiological data was recorded at a sample rate of 2000 samples per second.

The scenarios were recorded with 3 video cameras (Sony handycam) from three different angles at HD resolution. The original recordings were stored, and the clips were further compressed using XVID codec at a resolution of 480*856 pixels and stored in an avi container. The compressed clips were used for annotation.

The recordings were performed using 2 PCs, each recording synchronized data from 1 Kinect, 2 microphones, 1 head-tracker and one NeXus sensor using our own software (see setup in Figure 1). This is done by applying a global time stamp when the data are retrieved. The synchronization between the two PCs and the video cameras is done using two external monitors displaying the time stamp from each PC and that are visible by the cameras.

TABLE 1. SCENARIOS DESCRIPTION.

No.	Actor	Student
1	You have an important job interview and you are already running late so you rush to the bus. You don't have enough balance on your card and also forgot your wallet so you have no other means of payment with you. This interview is very important and this bus is the only bus that can ensure that you will be on time. Make sure you can still travel on this bus.	You work as a bus driver. There is a passenger who wants to travel but does not have sufficient balance on his/her card. Without a valid ticket you cannot allow a passenger to travel. If you let them ride without a valid ticket you risk a fine and can even lose your job.
2	It is in the middle of a cold winter night and you realize that you forgot your jacket in the night club you've just visited. You check your pockets but can't find the wardrobe token. You go back to the club, which is now closed, and ask the doorman to let you get your coat. Make sure you get your coat.	You work as a doorman at a nightclub. A common problem is that often coats are stolen, although there is a guarded cloakroom. You can only deliver coats based on tokens. The nightclub is currently closed and you cannot let people in anymore.
3	You are an employee of the Education Desk. Just before the start of a new quarter commonly many students who have forgotten to register on time come to ask for enrollment in courses. You are not allowed to register any more students since the deadline passed.	You are at the student desk and are late to register for a course. If you do not pass this course, you have to redo it next year and this causes delay in your graduation. Try to arrange that the employee at the counter can still register you for the course.

4. Annotation

4.1. Procedure and annotated dimensions

As a first step the recordings were manually segmented into utterances. This was done based on turn-taking and by splitting longer utterances at pauses. As the actors and the students had lively arguments during the role plays, the high amount of disagreement led to a high proportion of overlapping speech. Speaker identification and markers for overlapping speech are provided. The annotations were performed using the annotation tool Anvil [15].

The annotated dimensions are aggression, fear, valence, arousal, dominance. Except for valence, all dimensions were annotated on a 5 point scale. In the case of valence we expected the rating will mostly fall on the negative side of the scale so we chose a 9 point scale (-4 to 4) such that we obtain a 5 points granularity for the negative axis.

Each dimension was annotated by 3 raters (15 raters in total for the 5 dimension). Raters are students of a technical university. They received training, with detailed written definition and instructions, and they did a training session. Each rater was annotating only one dimension, giving separate scores for the actor and the student (in total 15 persons participated in annotation). It was emphasized that they had to pay special attention to all visible and audible signs that related to the dimension they are annotating (multimodal annotation).

4.2. Annotation outcome

The segmentation resulted in 2240 utterances. The proportion of speech for actors, students and overlapping speech is shown in Figure 2-a).

Inter-rater agreement in terms of Krippendorff's alpha [16] for each annotated dimension is shown in Table 2. Before applying this measure, the scores were z-normalized to account for personal rating biases [23]. The alpha was computed separately for the actors, the students, for actors on segments were the actors were speaking, for students on segments when the students were speaking, for actors when the actors were silent, for students when they were silent, for all segments that include speech (either actor or student) and finally for segments where there was silence.

The agreement for rating the actors is consistently higher than for rating the students for each annotated dimension. Nevertheless, the behavior of the actors and the students was rated for each dimension by the same raters. Hence, it cannot be concluded that the participants rating the student's behavior showed little task involvement, since they could achieve high agreement for rating the actor. What can be noticed from the data, and is probably the main reason for this discrepancy, is that the behavior of the actors was much more overt (e.g. ample gesturing, loud voice), while most students had a much subtler behavior. In the case of fear, all raters consistently scored level 1 (no fear) for the actors' behavior. While this intuitively signifies complete agreement, Krippendorff's alpha cannot be computed since it assumes variation in the data. For students, the alpha value scored 0.07, so we conclude that the annotation for fear is not usable. In addition the results show that the inter-rater agreement was higher for segments that include speech and lower for segments that include silence. This was the case both when considering the actors and the students. This finding indicates that listing non-verbal behavior is harder to interpret.

Figures 2(b-e) display the label distributions for aggression, fear, valence, arousal and dominance. As expected from the given roles, Figure 2-b) indicates that the actors were perceived more aggressive than the students. Actors were also rated with higher arousal scores, unanimously no fear, more dominant and with increased negative valence. The assumption that the data would fall on the negative side of valence is confirmed in Figure 2-c) - only 3.6% of the valence scores were positive.

5. Discussion

The collection of the NAA dataset was motivated by our final application of developing a Virtual Reality Ag-

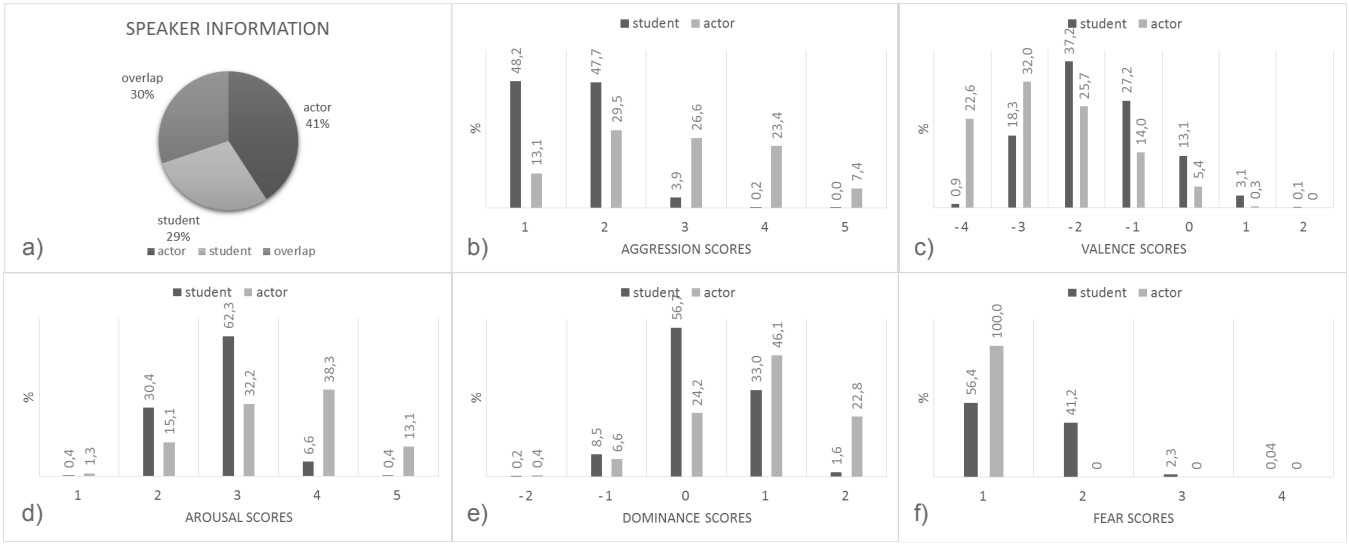

Figure 2. Distribution of mean annotated aggression scores (1 = low and 5 = high) for actors and students.

TABLE 2. INTER RATER AGREEMENT (KRIPPENDORFF'S ALPHA) FOR RATING AGGRESSION (AGG), VALENCE (V), AROUSAL (A), AND DOMINANCE (D) FOR ALL ACTORS SEGMENT(ACT), ALL STUDENT SEGMENTS (ST), FOR THE ACTORS ON SEGMENTS WHERE THE ACTOR IS SPEAKING (ACT_SP), FOR STUDENTS ON SEGMENTS WHERE THE STUDENT IS SPEAKING (ST_SP), FOR ACTORS ON SEGMENTS WHERE THE ACTOR IS LISTENING (ACT_LST), FOR STUDENTS ON SEGMENTS WHEN THEY ARE LISTENING (ST_LST), FOR ALL SPEECH SEGMENTS (SP) AND FOR ALL LISTENING SEGMENTS (LST).

	Act	**St**	**Act_sp**	**St_sp**	**Act_lst**	**St_lst**	**Sp**	**Lst**
Agg	.78	.27	.80	.33	.68	.11	.78	.49
V	.76	.60	.75	.60	.75	.60	.74	.65
A	.72	.31	.72	.34	.69	.24	.72	.52
D	.62	.45	.63	.49	.52	.31	.62	.40

gression Prevention Training (VRAPT) system. Ideally the dataset would contain recordings of forensic psychiatric inpatients in real situations that challenge their abilities for aggression management. However, controlled exposure of patients to real-life provocations in forensic clinics is not allowed. Recordings during aggression regulation therapy focus mostly on theory and a few role plays but were restricted for ethical concerns.

Our approach was to record role-plays between aggression training actors and naive participants. Role-plays are typically used in aggression regulation trainings and they will be part of VRAPT, which makes them a suitable option. Furthermore, during role-plays interactions evolve naturally, as would be the case in a real situation. No scripts were provided and the participants react to each other's behavior.

An important requirement was to have appropriate emotional / aggressive content. This was achieved by the scenarios' design and the use of professional aggression training actors. Advantages of using this special type of actors are the facts that they are able to easily escalate and deescalate a situation, and are proficient in reading the emotional state of the person they interact with. They are familiar with different techniques of putting the opponent in an uncomfortable situation, such as intimidation. In terms of aggression types, by watching the clips it can be observed that the data contains mostly instances of reactive aggression.

Regarding the realism of the dataset, the disadvantage of using actors is that they can be overacting. The naive participants on the other hand, do not share the danger of overacting. They are trying to handle the situation they are in the best they can, and to respond to the challenging behavior of the actors. As in the case of any type of recording (unless cameras are hidden), the presence of recording hardware can have an effect on behavior.

The responses from the surveys give more insight into how both actors and students perceived the interactions. All students replied that they found the actors and role-plays realistic and that they experienced physical arousal as the actors got really angry and were convincing. One student stated that the actors were over-reacting at some point. One student mentioned that at some point it was hard to stick to the requested goal (don't let the person travel without ticket) because in reality she would have done the opposite. The actors mentioned that they did their play according to the opponent and found it easier to interact with students that were more extrovert.

The use of physiological sensors can be a source of discomfort and might restrict natural movement. From responses to the survey we learn that 2 of the actors felt restricted by the electrodes, but none of the students did.

Emotional expressions result as a combination of push (physiologically driven) and pull (social regulation and strategic intention) factors [26]. The way in which the NAA dataset is constructed facilitates the study of both factor types. For the actors, pull factors have a dominant role. Their nonverbal behavior can be used as training material for overt behavior recognition. For the students, emotional

arousal is more genuine. Push factors can be studied based on physiological data while pull factors affect the final observed behavior, which can be used for overt behavior recognition. Interestingly, we observed cases of students displaying dominant and confident behavior accompanied by high heart rates, which shows that these factors interact.

The histograms depicting label distributions in Figures 2 show that a wide range of aggression levels and valence, arousal and dominance values appear. For fear the range of annotated values is limited: no fear (label 1) was constantly assigned for the actors while mild fear (level 2) appeared in 41% of the students' instance. Annotated labels are unbalanced for most dimensions which results from trading off a higher degree of naturalism for lack of control on the emotional / aggressive outcome.

While high inter-rater agreement was achieved for rating the actors, the agreement was low for rating the students. Furthermore, higher inter-rater agreement was observed for speaking behavior than for listening behavior. By watching the clips we observe that the actors' behavior is much more overt compared to the spontaneous reactions of students. All 4 actors use ample body language. For students this differs per person. In general the body language of the students is less prominent, but interesting signs of discomfort are visible (e.g. fidgeting). Less visible behavior is more challenging to rate, which was confirmed by our project partners with backgrounds in psychiatry and psychology who also viewed the clips. Low inter-rater agreement was experienced by other researchers too. In [28] examples of agreement rates for a set of databases and some of the problems are discussed. In general, acted datasets with clear emotion classes achieve high agreement, while datasets with more blended emotions and less overt behavior get lower scores.

The NAA dataset contains human-human interactions. However, in the final VRAPT system patients will interact with avatars which may have an impact on how they respond. Regarding this issue, research has shown that virtual characters can affect people's beliefs, behavior and can develop feelings of arousal and anxiety [24]. Therefore we expect that the avatars in VRAPT will be able to elicit negative arousal in patients. Further, from [10] where participants interacted with avatars still showed significant amounts of nonverbal behavior we expect that participants in VRAPT will behave similarly to when interacting with humans.

6. Possible research applications

Besides our envisioned application, in this subsection we give an account of research directions that could be approached using the NAA dataset.

The range of sensors used offers the possibility of studying aggression as well as negative affect in a multimodal way. Multimodal fusion is a complex and unsolved problem in part caused by how verbal and nonverbal signals arise and combine [19]. The relation and synchrony between different body manifestations can be studied using NAA. Compared to audio-visual datasets, NAA has the advantage of motion capture which facilitates such studies.

Even though facial expressions received a lot of attention, general posture and movement of the head is less studied [13]. The use of head trackers in NAA facilitates such studies, and recognition outcomes can be easily integrated in VR applications where HMDs are used.

The utterance based segmentation together with the recorded modalities offers the possibility of studying discourse and back channeling. Overlapping speech is usually discarded from the speech datasets. However, interruptions and overlaps as well as the amount of speech have a relation with dominance [14] and could be used in relation to aggression recognition given the annotations of overlapping speech in NAA. Moreover, the dataset enables studies of nonverbal behavior while listening and speaking, since annotations of both participants are available at all times.

The physiological signals can be useful to explore best signals to be used for aggression and negative affect recognition. They can also be used as proof of emotional arousal during student-actor interactions, and to investigate how push and pull factors combine for emotional expression.

Last but not least, the importance of incorporating context is increasingly acknowledged [21]. In [8] the role of several types of context is emphasized. Semantic context refers to the semantic content of words. It can be studied using NAA since the participants conduct a real meaningful conversation. The wide range of recorded modalities facilitates research on intermodal context, and the use of role-plays which enable interactions that evolve over time offers the possibility of studying temporal context as well.

7. Conclusion and future work

We introduced the negative affect and aggression (NAA) corpus consisting of interactions between professional aggression training actors and naive participants (students). The recordings contain audio, video, motion capture and physiological data. The final aim is to develop recognition software for Virtual Reality Aggression Regulation Training. However, the design, structure and content of the recordings support addressing a wide range of research questions.

Advantages of using aggression training actors are that they can easily escalate and deescalate, they are skilled in interpreting nonverbal behavior and construct the interaction in a way that matches their interest. The disadvantage is that they can be overacting. The naive participants do not share the danger of overacting, and respond naturally to the situation created by the scenario and the actors.

The use of unscripted role-plays facilitates interactions to build up naturally. They are similar to the final application of VR aggression prevention training. They also facilitate studying behavior in semantic and temporal context, while the wide range of recorded modalities support the use of intramodal context and work on multimodal fusion.

The data was segmented into utterances and annotated by multiple raters per each participant in interaction. The annotated dimensions are aggression, fear, valence, arousal and dominance, each on an ordinal scale. We observe a high inter-rater agreement for rating the actors, and a lower one

for rating the naive participants. This could be caused by the behavioral differences between the two categories: overt behavior for the actors and more subtle one for the students. The annotation shows that the recordings contain a wide range of aggression, valence, arousal and dominance levels. In addition, inter-rater agreement was higher when considering segments containing speech and lower for segments containing listening behavior.

In the near future we plan to further analyze the annotations and to use them for multimodal recognition. The recognition software will be integrated in the VR aggression prevention training, to give feedback to the therapist and to influence the avatar's behavior. The first version of the training system will be used in a pilot with forensic psychiatric inpatients. We will try to make more recordings during piloting such that we can compare the data of real patients with our original recordings and the match expectations.

References

[1] Batliner, A., Steidl, S., Nöth, E.: Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In: Proc. Satellite Workshop of LREC, pp. 28–31 (2008)

[2] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Interspeech, vol. 5, pp. 1517–1520 (2005)

[3] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S.: IEMOCAP: interactive emotional dyadic motion capture database. LREC **42**(4), 335–359 (2008)

[4] Chaquet, J., Carmona, E., Fernandez-Caballero, A.: A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding **117**(6), 633 – 659 (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. K. Scherer, T. Bänziger et E. Roach, éditeurs, A Blueprint for Affective Computing: a Sourcebook and Manual pp. 151–165 (2010)

[6] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al.: SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In: Proc. Int. Conference on Autonomous Agents and Multi-agent Systems, pp. 1061–1068 (2014)

[7] Dodge, K.A., Coie, J.D.: Social-information-processing factors in reactive and proactive aggression in children's peer groups. Journal of personality and social psychology **53**(6), 1146 (1987)

[8] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a new generation of databases. Speech communication **40**(1), 33–60 (2003)

[9] Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica **1**, 49–98 (1969)

[10] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., Devault, D., Marsella, S., Traum, D., Rizzo, A.S., Morency", L.P.: The Distress Analysis Interview Corpus of Human and Computer Interviews. In: Proc. LREC (2014)

[11] Grimm, M., Kroschel, K., Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database. In: Multimedia and Expo, 2008 IEEE International Conference on, pp. 865–868. IEEE (2008)

[12] Gunes, H., Piccardi, M., Pantic, M.: From the lab to the real world: affect recognition using multiple cues and modalities. In: J. Or (ed.) Affective computing: focus on emotion expression, synthesis, and recognition, pp. 185–218 (2008)

[13] Hammal, Z., Cohn, J.F., Heike, C., Speltz, M.L.: What can head and facial movements convey about positive and negative affect? In: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, pp. 281–287 (2015)

[14] Itakura, H.: Describing conversational dominance. Journal of Pragmatics **33**(12), 1859–1880 (2001)

[15] Kipp, M.: Anvil - a generic annotation tool for multimodal dialogue. In: Proc. 7th European Conference on Speech Communication and Technology (Eurospeech) (2001)

[16] Krippendorff, K.: Computing Krippendorff's alpha reliability. Departmental papers (ASC) p. 43 (2007)

[17] Kunst, M.J., Bogaerts, S., Winkel, F.W.: Peer and inmate aggression, type d-personality and post-traumatic stress among dutch prison workers. Stress and health **25**(5), 387–395 (2009)

[18] Lefter, I., Burghouts, G., Rothkrantz, L.: An audio-visual dataset of human-human interactions in stressful situations. Journal on Multimodal User Interfaces **8**(1), 29–41 (2014)

[19] Lefter, I., Burghouts, G., Rothkrantz, L.: Recognizing stress using semantics and modulation of speech and gestures. IEEE Transactions on Affective Computing (99), 1–1 (2015)

[20] Lefter, I., Nefs, H., Jonker, C., Rothkrantz, L.: Cross-corpus analysis for acoustic recognition of negative interactions. In: Affective Computing and Intelligent Interaction, pp. 132–138 (2015)

[21] Lefter, I., Rothkrantz, L., Burghouts, G.: A comparative study on automatic audiovisual fusion for aggression detection using meta-information. Pattern Recognition Letters **34**(15), 1953 – 1963 (2013)

[22] 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. Affective Computing, IEEE Transactions on **3**(1), 5–17 (2012)

[23] Metallinou, A., Katsamanis, A., Narayanan, S.: Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. Image and Vision Computing **31**(2), 137–152 (2013)

[24] Qu, C., Ling, Y., Heynderickx, I., Brinkman, W.P.: Virtual Bystanders in a Language Lesson: Examining the Effect of Social Evaluation, Vicarious Experience, Cognitive Consistency and Praising on Students' Beliefs, Self-Efficacy and Anxiety in a Virtual Reality Environment. PLoS ONE **10**(4) (2015)

[25] Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Automatic Face and Gesture Recognition (FG), pp. 1–8. IEEE (2013)

[26] Scherer, K.R., Bänziger, T.: On the use of actor portrayals in research on emotional expression. In: K.R. Scherer, T. Bänziger, E.B. Roesch (eds.) Blueprint for affective computing: A sourcebook, pp. 166–176. Oxford, England: Oxford university Press (2010)

[27] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al.: The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In: proc. INTERSPEECH. Citeseer (2013)

[28] Siegert, I., Böck, R., Wendemuth, A.: Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. Journal on Multimodal User Interfaces **8**(1), 17–28 (2013)

[29] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proc. 3rd ACM international workshop on Audio/visual emotion challenge, pp. 3–10. ACM (2013)

[30] Huis in t Veld, E.M., Van Boxtel, G.J., de Gelder, B.: The Body Action Coding System I: Muscle activations during the perception and expression of emotion. Social neuroscience **9**(3), 249–264 (2014)

[31] Vinciarelli, A., Dielmann, A., Favre, S., Salamin, H.: Canal9: A database of political debates for analysis of social interactions. In: Affective Computing and Intelligent Interaction, pp. 1–4. IEEE (2009)