The Recognition of Acted Interpersonal Stance in Police Interrogations

Sophie Spitters
Behavioral Sciences
University of Twente
Enschede, Netherlands
s.j.i.m.spitters@student.utwente.nl

Merijn Sanders
Industrial Design Engineering
University of Twente
Enschede, Netherlands
g.m.a.sanders@student.utwente.nl

Rieks op den Akker
Merijn Bruijnes
Human Media Interaction
University of Twente
h.j.a.opdenakker|m.bruijnes@utwente.nl

***Abstract*—This research aims at finding how suspects in police interrogations express their interpersonal stance -in terms of T.Leary's interpersonal circumplex- through body postures and facial expressions and how this can be simulated by virtual humans. Therefore, four types of stances were acted by eight actors. To see if the resulted postures are valid, short recordings were shown online in a survey to subjects who were asked to describe them by a selection of a number of adjectives. Results of this annotation task show that some stance types are better recognized than others. Validity (recognizing the intended stance) and inter-rater agreement do not always go hand in hand. The body postures and facial expressions of the best recognized fragments are annotated so they can be implemented in the artificial agent. The results of this study are used in a serious game for police interrogation training where the role of the suspect is played by an artificial embodied conversational agent.**

I. Introduction

An important quality of social interaction is that people recognize the emotions of one another. It is well known that in conversations emotions are not only expressed by the verbs that are spoken but also by other "channels of communication". It is said that "no more than 30 to 35 percent of the social meaning of a conversation or an interaction is carried by the words", [1]. The other part is expressed through facial expressions, gestures, postures and prosody. The ability to recognize interpersonal stance, a type of affect that is focused on in this research, makes a good part of the conversational skills that in many professions are required. This research is carried out with a focus on police interrogations with the aim to build artificial embodied conversational characters. These characters have to play the role of a suspect in a serious game by means of which police trainees learn to interview witnesses or interrogate suspects. Trainees learn to see how the behavior of a suspect is related to their own behavior. Interpersonal stance is a core construct in training the interrogation skills of Dutch police trainees. T. Leary's theory of interpersonal relations is used as a framework to analyze interpersonal stance. Leary's model, the interpersonal circumplex also known as *Leary's Rose* ([2]) is presented by a circular ordering of eight categories of interpersonal behavior, situated in a two-dimensional space spanned by two orthogonal axes. The horizontal axis is affiliation (positive versus hostile), the vertical one is the power axis (dominant versus submissive) (Figure 1). Research demonstrated the value of the model for integrating a broad range of psychological topics, [3], [4].

If an artificial suspect character is used, it must be sure that he expresses the various stances and emotions in such a way that it is convincing and recognized by the police trainee. Therefore, the main goal of this research is to relate body postures and facial expressions to the interpersonal stances expressed in Leary's Rose. This study involves the generation of the expression of interpersonal stances as well as the recognition of stances by independent observers. The question is: "are there typical postures or facial expressions that express a particular interpersonal stance towards the interrogator?" To answer this question, actors were asked to depict all stances. The resulting fragments were then annotated and analyzed. Are the depicted stances reliable? That is: "are people able to recognize the interpersonal stances expressed by actors from short video fragments?" and "do different observers agree"? Body language comes in clusters of signals and postures, depending on the internal emotions and mental states. Recognizing a whole cluster is thus far more reliable than trying to interpret individual elements [5]. An annotation task was carried out where observers had to label whole video fragments showing acted stances. From a list of adjectives that people use to describe the stances of Leary's Rose (cf. [4]) observers had to select a number of adjectives that they thought best describes the stances being depicted in the videos.

II. Method: Generating and Annotating Stances

The method followed in this research consists of two parts. First of all, clips of the interpersonal stances were generated by the use of actors. Secondly, the validity of these depicted stances was assessed by means of annotation.

A. Generating Interpersonal Stances

The clips of the interpersonal stances were generated by using actors. Eight actors have taken part in this experiment. Four of them were members of a theatre club and thus, had some acting experience. The rest were novices. Each actor had to depict four stances. The stances correspond to the quarter segments of the rose in Figure 1. These four segments are abbreviated as 'DP'(Dominant-positive), 'SP'(Submissive-Positive), 'SH'(Submissive-Hostile), 'DH'(Dominant-Hostile), as can be seen in the figure.

All actors were given the same scenario. They had to imagine they were suspected of shoplifting and in the middle of an interrogation. Then, they watched a computer screen where a video fragment shows a police interrogator addressing them and asking them what happened. The actors were then asked to give a short response (max. $10s$) with a stance. This produced 32 video recordings that were used in the survey.

However, the actors differed in the instructions they got on how to depict interpersonal stances. Half of the actors were selected to the 'theory-condition' and the other half to the 'scenario-condition'. The actors in the 'theory-condition' got theoretical instructions about Leary's rose, [2]. To help them get an even more concrete idea of what the stances mean, several adjectives that capture the meaning of the stances were given. This was a random selection from the list created by Rouckhout and Schacht [4]. The summary of their instructions were captured in an image, that is shown in Figure 1. The actors in the 'scenario-condition' got a specific scenario for each stance, that was directly linked to the interrogation setting and to the question of the interrogator. The scenario is supposed to provoke a reaction in a certain stance in a more natural way than is the case in the 'theory-condition', as the workload of processing and interpreting the theory is reduced and actors can put their resources into entering into the part they are playing.

B. Annotating Interpersonal Stances

An online survey was created to see if the video recordings were valid. A convenience sample was used that consisted largely of students. The subjects were asked to annotate 8 fragments of a total of 64, 32 with sound and the same 32 fragments without sound. The distinction between with and without audio was used to check if people are better at recognizing interpersonal stances if they have more information (sound). The video fragments were assigned randomly to the subjects, but in such a way that a subject viewed exactly one clip of each actor. For annotating the fragments, a semi-forced format was used, meaning that subjects were given a list of 32 adjectives and were free to select any number of adjectives (with a minimum of four) that they thought fit the stances expressed in the fragments. For a discussion about formats refer to Busso et al.[6]. The list of adjectives was the same as used in the theoretical instruction for the actors. Note that annotators did not know the stance categories of the adjectives. The adjectives were given in a random order.

III. Results

In order to reach the main goal of this research, describing postures associated with interpersonal stance, it needs to be investigated if the acted stances are valid. Therefore, it is tested if people recognize the acted stances. First, the distributions of the responses are focused on to get a first indication of how well people perform at annotating the videos. Second, the individual judgements are investigated to see how well individuals recognize the stances and to see the extent to which individuals agree with each other on what stance they think is being depicted in a video. Finally, the best recognized videos for each stance are annotated to extract key poses and gestures that can be used in a conversational agent.

A. Responses

1) Distribution of responses: To get a first indication of how well acted stances are being recognized, it is tested if adjectives belonging to the depicted stance are more often chosen than adjectives from different stances. To adjust for respondents choosing many adjectives to annotate a fragment and therefore having a bigger influence, calculations have been made for each annotation reporting the percentage of adjectives that belong to the different stance categories. The distributions of these percentages are used in this section.

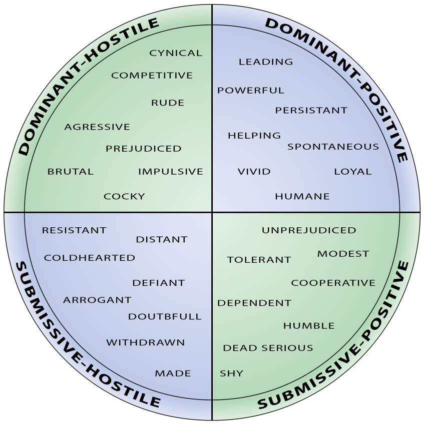

Fig. 1. Summary of the instructions given to the actors in the 'theory-condition' in order to express interpersonal stances(translated from dutch)

For each of the four stances that are depicted by the actors, a pie chart has been made that shows the mean percentages of annotated adjectives belonging to each stance-category, see Figure 2. The figure gives a first indication of how good respondents are at recognizing the depicted stances. It is striking that stance category 'SH' seems to be chosen the most by the respondents independent of what stance the actor depicted.

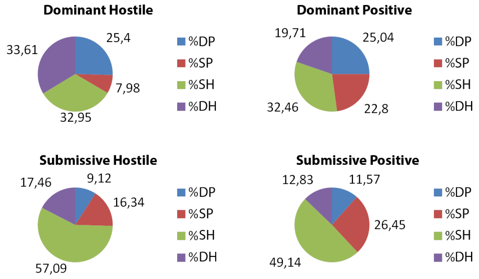

Fig. 2. For each of the acted stances, the pie chart shows the mean percentages of chosen adjectives belonging to the four stance categories.

2) All judgements: In total 84 subjects each judged 8 fragments by selecting at least 4 adjectives from a list of 32 adjectives that they found most appropriately describing the stance acted by the actor shown in the fragment. Since the adjectives belong to one of 4 categories of Leary's circumplex, it is interesting to see how often there is a match between the category of the adjective chosen and the category of the stance acted out in the fragment. A judgement of a fragment by a subject is called:

- Perfect, when *all* the adjectives that a subject has chosen as describing that fragment belong to the same class as the stance that was supposed to be expressed by the actor in the fragment.

- Correct, when there is a unique category with a maximum number of adjectives selected and this category is the same as the class of the fragment.
- Semi-correct, when the category that has the maximum number of adjectives chosen is the same as the class of the fragment.
- Wrong, if it is not semi-correct.

Table I shows for each of the categories how many times the judgments were perfect, correct, etc. The total number is 672. There are small differences in the numbers of fragments in each of the four categories. From the total of 672 judgments 162 judgments concern DP fragments, 178 concern DH fragments, 157 concern DP and 175 SH fragments.

Table II shows the confusion matrix. It shows for each of the stances (rows) how often fragments of that stance were assigned the four classes if we take the stance category with the maximum number of adjectives as the stance assigned. In case there is no unique stance category with a majority then the decision is X (undecided). From the numbers in Table II we compute the precision, recall and F-values (Table III).

The SH and DH (opposition) categories have clearly higher F-measures than the two together categories DP and SP. The highest precision is obtained for class DH.

	Judgments				
CAT	PERF	CORR	SEMICOR	Wrong	Total
DP	3	28	53	109	162
SP	1	27	47	110	157
SH	22	113	138	37	175
DH	0	52	84	94	178

TABLE I. THE COUNTS HOW MANY TIMES SUBJECTS ASSIGNED THE "CORRECT" STANCE TO THE FRAGMENTS IN EACH OF THE FOUR CATEGORIES. FOR EXPLANATION OF WHAT "CORRECT" MEANS SEE THE MAIN TEXT.

	Chosen Stance				
CAT	DP-C	SP-C	SH-C	DH-C	X-C
DP	28	31	43	20	40
SP	13	27	77	7	33
SH	6	15	113	14	27
DH	36	6	49	52	35

TABLE II. THE COUNTS HOW MANY TIMES SUBJECTS ASSIGNED STANCES TO THE FRAGMENTS IN EACH OF THE FOUR CATEGORIES.

CAT	Acc	Prec.	Recall	F
DP	0.72	0.34	0.17	0.24
SP	0.73	0.34	0.17	0.24
SH	0.66	0.40	0.65	0.52
DH	0.75	0.56	0.29	0.38

TABLE III. THE ACCURACY, PRECISION, RECALL AND F-VALUES FOR EACH OF THE FOUR STANCE CATEGORIES. THESE FIGURES ARE BASED ON THE FIGURES IN THE CONFUSION TABLE II

α - stance categories							
DP	-DP	SP	-SP	SH	-SH	DH	-DH
0.12	0.23	0.08	0.24	0.03	0.21	0.22	0.15

TABLE IV. THE α VALUES COMPUTED FOR THE FRAGMENTS OF EACH ACTED STANCE AND FOR ALL FRAGMENTS EXCLUDING THE FRAGMENTS OF A SPECIFIC ACTED STANCE USING KRIPPENDORFF'S METHOD WITH DICE METRICS FOR DISTANCES BETWEEN VALUES.

B. *Inter-Annotator Agreement*

Did subjects agree on the selection of adjectives for describing the different fragments? If subjects do not agree in their accounts of the stance taken by the actors in the fragments they have judged then it is difficult to say what the stance is that the actor takes. Thus, we analyzed the judgments for inter-annotator agreement. How do we measure this? If we look at the "coding task" at hand we see that it has the following properties.

- There is a large number of annotators (84).
- Not all annotators annotated all fragments.
- The label set used in the annotation task is large (32 adjectives).

Because of these properties we use Krippendorff's α agreement method for computing a reliability measure ([7], p.222). We apply the method for many observers, many nominal categories, and for missing values ([7], p.232). We could take as labels the subsets of adjectives. But the number of label subsets is too large. We defined the labels as sets of stances. After all we are interested in stances annotators assigned to the fragments and each adjective uniquely refers to one stance. The set C of stances is assigned to a fragment if C contains all and only those stance categories that have a maximum number of adjectives in the set of adjectives chosen by the annotator. We need a distance metrics on these sets to compare the labels assigned by two annotators. For comparing two set values we use a distance metrics based on the similarity measure on sets known as *Dice coefficient*:

$$sim(C_1, C_2) = \frac{2|C_1 \cap C_2|}{|C_1| + |C_2|} \tag{1}$$

($sim(\emptyset, \emptyset) = 1$) The distance between two sets C_1 and C_2 is:

$$\delta(C_1, C_2) = 1 - sim(C_1, C_2)$$

We use δ with sets of stance labels.

The α score for the whole corpus results in an value of 0.22. We also calculated α for parts of the corpus containing only fragments of a certain intended stance, see Table IV. This table also shows the α values for the corpus without the parts containing fragments of a specific stance. Remarkable are the exceptional values for the DH fragments. Remember that this is the class that also has the highest precision value. DH stance behavior is easier to recognize (and perform!) than the other types of stances.

C. Mute vs. Sound and Theory vs. Role-play

We had two different settings in which actors were asked to perform the four stances. Four of the eight actors were recorded in the Theory setting. The other four in the Role play setting. There where also 2 different settings in which the fragments where shown to the survey participants, namely with and without sound. In this section these conditions will be further explored to see how these influence the judgments.

In Figure 3 the judgments with sound, muted and total are visualized in a bar graph. The total value is divided by 2 which represents the judgments if sound and mute would be fully equally distributed. From this graph we can compare the judgments of the S- and the M-fragments and we see that there appear to be no significant differences between the fragments with and without audio. In Figure 4 the judgments for the theory setting, roleplay setting and total are visualised in a bar graph. The total value is divided by 2 which represents the judgments if Theory and role play would be fully equally distributed. As can be seen here it appears that overall the Theory settings induces acted stances that are better recognized than those in the Role play setting.

For these different settings α are also calculated the results are shown in Table V. It shows the α values for the whole class of fragments and for the class of S-fragments (with audio) and the class of M-fragments with muted audio. These values are low. There is slightly more agreement on the fragments with audio as there is on the fragments without audio. Clearly, the Theory play judgments have a higher inter-rater agreement.

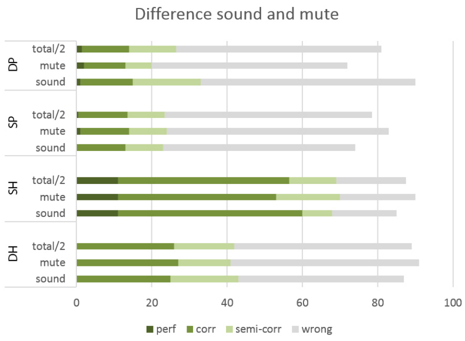

Fig. 3. Difference between mute and sound

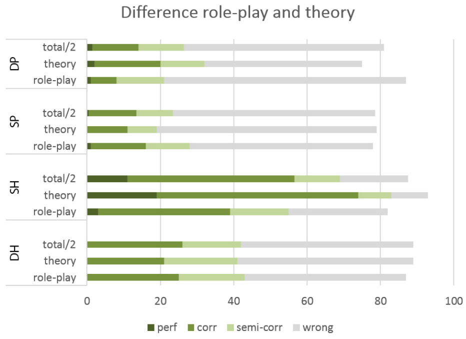

Fig. 4. Difference between theory and role-play groups

	α - audio		α - condition	
ALL	Mute	Sound	Role Play	Theory
0.22	0.21	0.23	0.15	0.27

TABLE V. THE α VALUES COMPUTED FOR FRAGMENTS WITH AND WITHOUT SOUND AND FOR THE ROLE PLAY AND THEORY FRAGMENTS USING KRIPPENDORFF'S METHOD WITH DICE METRICS FOR DISTANCES BETWEEN VALUES.

Actors in Theory-Condition							
T01		T02		T03		T04	
score	α	score	α	score	α	score	α
72	0.16	90	0.22	58	0.03	114	0.38
Actors in Role Play-Condition							
R01		R02		R03		R04	
score	α	score	α	score	α	score	α
61	0.21	68	0.24	62	0.04	43	0.01

TABLE VI. THE SCORES AND α RELIABILITY VALUES FOR EACH OF THE 8 ACTORS

D. Are some actors better than others?

Are some actors better than others in the sense that the stances they perform are easier to recognize by the subjects? We compute for each of the actors a score. For each of the actors we look at the 84 judgments in which the actor acted. If the judgment is perfect we add 3 points to the score, if it is correct we add 2 points to the score, if it was semi-correct we add 1 point to the score. The resulting scores are in Table VI. Actor T04 scores significantly higher than the mean score and actor R04 scores lower than the mean. What is the impact of these two actors on the $alpha$ values? If we remove all judgments with R04 α slightly raises from 0.22 to 0.25. If we remove T04 α becomes 0.19. If we only take the fragments with actor T04 α raises to 0.38. Our analysis confirms that some actors are better than others and that good acting has a significant impact on inter-annotator agreement.

Since fragments were assigned randomly to observers there is a chance that differences in scores are due to the observers not to the actors. In order to cancel this out we looked at those judgments by subjects that both annotated the same stance by the same actors. A paired t-test comparing scores for R04 and T04 on the 24 pairs of judgments of the same stances by the same subjects gives: $t(23) = 4.796$ ($p << 0.0001$). In all but one case the judgement of a subject has a higher score with actor R04 than with actor T04. In all other cases T04 scores equal (9 times) or higher (14) than R04. This gives sufficient evidence to rule out that the higher scores for T04 compared to those for R04 are due to the judges assigned to them. Table VI also contains the α reliability values for the 8 actors. Figure 5 shows the relation between scores and α values. It shows that the ratio between scores and α values varies considerably. For actors $R01$ and $R02$ they are much higher than for R03. The Spearman correlation between α and score equals 0.833 (significance $p < 0.05$, 2-tailed). Overall, there is a reasonable correlation between the inter-annotator agreements and the validity. But for some actors (e.g. R03) a higher score (validity: agreement between judgment and intended stance) goes with a low inter-annotator agreement and for others (e.g. R01, R02) a lower score (validity) goes with a higher inter-annotator agreement, i.e. relatively more annotators agree on the stance they see but it is not the stance as it was intended. We see that an actor being good has two different senses: he performs the stance that was asked to act, or he performs a stance that is recognized by a majority of observers. They are not independent, though: high validity means high agreement.

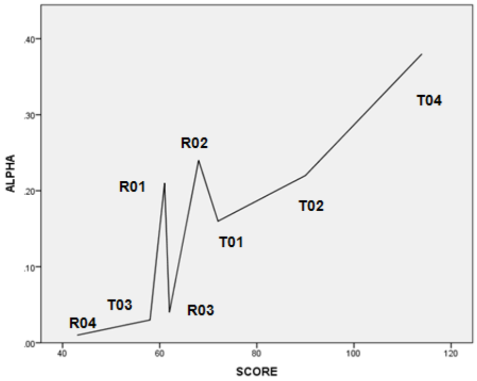

Fig. 5. Judgment scores and inter-annotator agreement for each of the 8 actors.

E. Best Fragments

As described before, the relation between depicted stance and perceived stance seems rather weak. This is why it could be difficult to clearly define a typical and valid posture that depicts a certain stance. Nevertheless we will try to qualitatively describe the best fragment of each depicted stance. In order to determine which fragments are the best, all 4 stances of all 8 actors, which represents all fragments, are judged and plotted. The plot showing scores versus α values is shown in Figure 6. For practical reasons the actors are numbered consecutively wherein actor number 1 till 4 represent the four actors in the theory-condition and 5 till 8 represent the four actors in the role-play condition. As can be seen in this plot α values are very low. This is caused by the small amount of respondents on each separate fragment. They will not be used in the selection of the fragments. The four fragments (one for each stance type) selected based on the best judgement scores are indicated by an arrow in the plot.

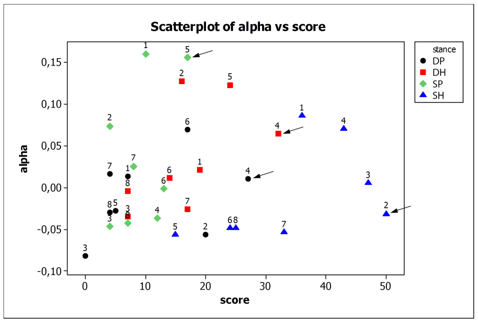

Fig. 6. Judgment scores and inter-rater agreement for each of the 8 fragments.

F. Comparing Depicted Stances

In order to be able to directly use the results in the police interrogation game, the best recognized fragments for each stance were annotated. For annotating body language, the body action and posture coding scheme (BAP)[8] is used. For annotating facial expressions, the facial action coding scheme (FACS) [9] is used. When for each stance category stills from the three best recognized fragments are compared, it stands out there are many similarities within one category and many differences between categories (see Figure 7). First, stills belonging to a dominant category participants look straight at the artificial agent and have an upright body posture. Differences between the Dominant Positive and the Dominant Hostile category are that the facial expressions in the positive category are more friendly, indicated by the combination of more smiling, wide open eyes and high eyebrows. The hostile category on the other hand is more aggressive as can be seen by the aggressive arm movements. Stills belonging to the submissive categories show more closed body postures and avoidance of eye contact. The body postures seen in the Submissive Positive category can be interpreted as shy. However, the postures in the Submissive Hostile category are more passive aggressive with the arms crossed and leaning backward. Their facial expressions match with lips that are curled downwards. These observations comply with the idea that dominant postures are used to intimidate or show that you are in control in contrast to submissive postures that are used when confrontation is avoided.

IV. Conclusion

A. Recognizing Depicted Interpersonal Stances

The results of the annotations show a clear correlation between the stances that are acted in the videos and the adjectives that are chosen in the judgments. However, there are many judgments in which subjects choose adjectives that belong to an other category than the stance category that was intended by the actor. The stance that is recognized best is ‘SH’ which can be expected while adjectives belonging to ‘SH’ are assigned most independent of the stance that is depicted. The rationale behind this may be that all but one of the most frequently used adjectives from ‘SH’ could be interpreted as an indication of the video fragments being acted. This makes sense, because all the interpersonal stances in the videos are indeed acted and several actors commented that the task felt unnatural to them what could have influenced the naturalness of their acting.

To see if the annotators agreed with each other inter-annotator agreement is calculated. With α of 0.22 the inter-annotator agreement is rather low. Earlier studies that look at inter-annotator agreement in a stance annotation task using Leary’s circumplex already showed that this is a hard task, (see ,e.g., [10]). In this experiment audio does not add necessary information for annotating interpersonal stance. It has to be noted that some videos contained silent acting. The judgements of actors in the theory-condition did seem to differ from the judgements in the role play-condition. For most stances, fragments with actors from the theory-condition seem to be better recognized. This is most obvious with the acted stance ‘SH’ where 19 of 22 perfect judgements are in the theory-fragments. It could be of influence that the actors in the

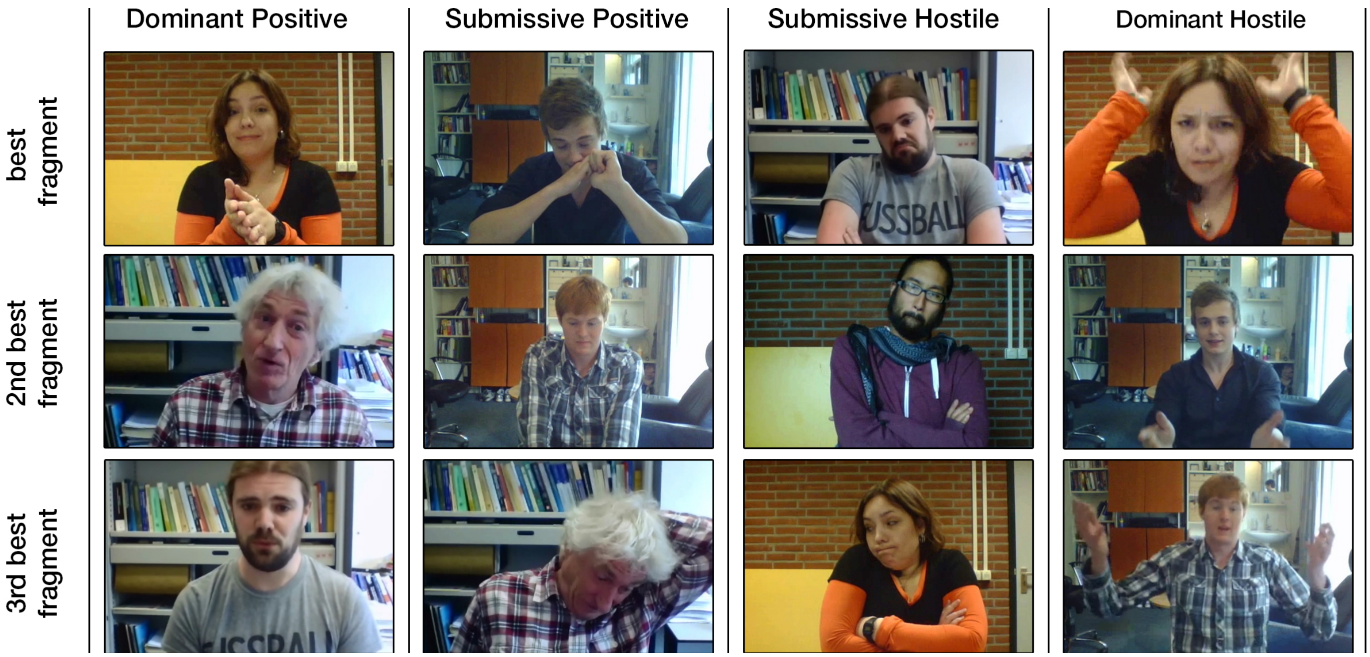

Fig. 7. Screen-shots of relevant postures of the fragments within each depicted stance category that were recognized best.

theory-condition had the exact same list of adjectives in their instructions as the list that was used in the survey. Some actors are better than others in the sense that they better put the stance they intended to show on stage, others are better in the sense that the stance they act is recognized by more spectators. The ratio between validity and inter-annotator agreement differs per actor.

B. Typical Expressions of Interpersonal Stances

In searching for typical postures and facial expressions that express the interpersonal stances the best best recognized fragments were investigated further. It was rather difficult to see if one fragment is better than the other, because there were so few annotations per fixed actor and stance. Still, the 'best' fragments were reviewed. It is striking to see that most fragments where the acting is exaggerated are recognized best. For making the virtual suspect this is okay while the interrogation game tries to get police trainees familiar with the effects of Leary's theory. It is more important that stances are recognizable than that stances are very realistic. Using the terminology of CogInfoCom [11] we see that for practical purposes in order to communicate the stance by means of a virtual character we transform the real stance by means of caricatural animations that exaggerate the characteristic behavioral elements we identified in this study. Furthermore, when stills from the best fragments are compared similarities within stance categories and differences between these categories are apparent. In summary, it can be seen that dominant postures are upright with a gaze straight at the conversational partner while submissive postures are more closed with a gaze away from the conversational partner. However, the most valuable lesson learned from this experiment is that it is hard to act a stance and -maybe even more valuable- that observers often see diverse aspects in the behaviour of someone. People apparently often show a mixture of stances.

ACKNOWLEDGMENT

This publication is supported by the Dutch national program COMMIT.

REFERENCES

[1] R. Birdwhistell, *Kinesics and Context*. University of Pennsylvania Press, 1970.

[2] T. Leary, *Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation*. New York: Ronald Press, 1957.

[3] J. Wiggins, P. Trapnell, and N. Phillips, "Psychometric and geometric characteristics of the revised interpersonal adjective scales (ias-r)," *Multivariate Behavioral Research*, vol. 23(4), pp. 517–530, 1988.

[4] D. Rouckhout and R. Schacht, "Ontwikkeling van een nederlandstalig interpersoonlijk circumplex," *Diagnostiekwijzer*, vol. 4, pp. 96–118, 2000.

[5] A. Mehrabian, *Nonverbal communication*. Transaction Publishers, 2009.

[6] C. Busso and S. S. Narayanan, "Recording audio-visual emotional databases from actors: A closer look," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008, pp. 17–22.

[7] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*. SAGE Publications, Second Edition, 2004.

[8] N. Dael, M. Mortillaro, and K. R. Scherer, "The body action and posture coding system (bap): Development and reliability," *Journal of Nonverbal Behavior*, pp. 1–25, 2012.

[9] P. Ekman and W. V. Friesen, *Facial action coding system: A technique for the measurement of facial movement*. SAGE Publications, Second Edition, 1978.

[10] F. Vaassen and W. Daelemans, "Emotion classification in a serious game for training communication skills," in *Computational Linguistics in the Netherlands 2010: selected papers from the 20th CLIN meeting*. LOT, 2010.

[11] P. Baranyi and A. Csapo, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, p. 6783, 2012.