Inferring a User's Intent on Joining or Passing by Social Groups

Andrea Bönsch
Visual Computing Institute,
RWTH Aachen University, Germany
boensch@vr.rwth-aachen.de

Alexander R. Bluhm
Department of Computer Science,
RWTH Aachen University, Germany

Jonathan Ehret
Visual Computing Institute,
RWTH Aachen University, Germany

Torsten W. Kuhlen
Visual Computing Institute,
RWTH Aachen University, Germany

Figure 1: Modeling user-awareness: Based on social cues of the user, our classification scheme infers her intent on either joining or passing-by free-standing, conversational groups, triggering an appropriate reaction of the individual group members.

ABSTRACT

Modeling the interactions between users and social groups of virtual agents (VAs) is vital in many virtual-reality-based applications. However, only little research on group encounters has been conducted yet. We intend to close this gap by focusing on the distinction between joining and passing-by a group. To enhance the interactive capacity of VAs in these situations, knowing the user's objective is required to show reasonable reactions. To this end, we propose a classification scheme which infers the user's intent based on social cues such as proxemics, gazing and orientation, followed by triggering believable, non-verbal actions on the VAs. We tested our approach in a pilot study with overall promising results and discuss possible improvements for further studies.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; **User studies**.

KEYWORDS

virtual agents, social groups, joining a group, virtual reality

ACM Reference Format:
Andrea Bönsch, Alexander R. Bluhm, Jonathan Ehret, and Torsten W. Kuhlen. 2020. Inferring a User's Intent on Joining or Passing by Social Groups. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423862

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

1 INTRODUCTION

Embodied, computer-controlled and anthropomorphic virtual agents (VAs) in virtual reality (VR) function primarily as interaction partners for users (e.g., [22]) and as background characters enlivening the virtual scenarios (e.g., [38]). To this end, VAs in a shared space require an *interactive capacity* allowing them to be autonomous, intelligent, conversational, and life-like. By this means, they can show plausible reactive and proactive behavior towards other VAs as well as the user.

In this work, we focus on encounters between a user and groups of VAs. Observations in real-life showed, that about 70% of humans [25] tend to form so called social groups [35]. Thus, they engage in a relationship by means of a shared action [15] in social environments. To improve the VR experience in enlivened scenarios, this human behavior can be simulated with VAs by grouping them into interactive groups of different sizes. However, this rises the need for an *interaction concept* between the user and such virtual, social groups. Basic components of such a concept are fundamental user-group interactions such as joining (and leaving) a social group, or passing-by. Upon joining, direct user-agent interactions are required, e.g., engaging in the joint action such as a conversation.

In our work, the VAs remain in their positions, meaning they are not moving around and groups do not split or merge. Thus, we focus on the basic components of user-group interactions in a *stationary*

context[1]. In this context, social groups are commonly free-standing, conversational groups [1] as known from school yards or various social gatherings. Embedding a user thus requires two interactive capacities for the VAs: First, they need to infer the user's intent based on his or her social cues: while approaching, a user might want to simply pass-by to reach another location in a scene or the user wants to join a group to listen or to actively participate in a conversation (cp., e.g., [27] for single VAs). Second, the VAs need to react adequately to the inferred intent. This includes expressing user-awareness, e.g., by means of gazing or greetings, as well as showing natural social behavior, e.g., by respecting a user's personal space [16].

To the best of our knowledge previous research focused on agent-group encounters (e.g., [19, 40]) or on a user approaching single VAs [27] or dyads [33]. Thus, our contribution is a classification scheme for encounters of a user and social groups in a shared, immersive scene within a stationary context. We thereby exclude the direct, e.g., verbal user-agent interaction while focusing purely on the identification of join- and pass-by-intents with consecutive, adequate reactions of the group members.

The paper's remainder is structured as follows: After providing related work on social groups w.r.t. group formations and interactions in Sec. 2, we present our classification scheme to infer the user's intent in Sec. 3. A pilot study evaluating the scheme is detailed in Sec. 4, the insights gained are discussed in Sec. 5 and summarized in Sec. 6.

2 RELATED WORK

Designing VAs for natural, social interactions in VR is non-trivial. Different behavior patterns for plausible actions are required to give users the illusion of lively and human-like interaction partners. One basic module is the VA's awareness of its surrounding focusing on perceiving agents and users being there. As indicated by Bönsch et al., this awareness needs to be expressed explicitly, e.g., in terms of gazing patterns and proxemics [7]. Especially proxemics, defined as the usage of space during social interactions, is of prime importance. Research has proven, that VAs need to respect the personal space [16] of other characters [13] as well as of the user [2]. Thus, the interpersonal distance is one aspect influencing the arrangement of interactants in a given scene. Thereby, distance perception in VR [14] and the availability of a body avatar as permanent reference frame to become more aware of the surrounding environment [24] have to be considered. A second aspect is the need of a shared space between the interactants, to which all have direct access [20]. The respective orientational and positional arrangement is referred to as F-Formations [12, 23, 31, 36]. While different natural arrangements for a dyad exist, social groups of three or more interactants typically form a circular arrangement [20] as presented schematically in Figure 3(a). Thus, simulating social groups in VR requires a natural positioning behavior of all group members [31]. Research by Ennis et al. indicated that a circular shape with interpersonal distances between .4m and 2.1m, depending on the VAs' relationship, is a good initial group shape for a group of virtual conversers [13]. Throughout the social interaction the group shape then needs to be adapted and refined, as stated by Jan et al. [19]. They propose a social force model attracting individual group members towards the current speaker while using a repelling force to maintain an appropriate interpersonal distance between all group members. Furthermore, they account for group size increases or decreases while also allowing for situation-dependent group splits and merges. In contrast, Pedica et al. propose a reactive framework based on the territorial behavior of humans to model group dynamics during social gatherings [31].

Upon the aforementioned proxemics and group shape, it is necessary to model the joint interaction in a plausible and natural way. Commonly, the social groups in VR-based scenarios are conversational, hence a natural course of the discussion needs to be simulated. This involves not only gazing patterns or appropriate mimics, but also aspects such as interruptions and turn-taking [13, 18, 32]. Furthermore, an adequate welcome for new group members is required, e.g., a head node [11], mutual gaze, smiling, and adapted proxemics [10]. Especially if users are supposed to interact more often with the group members, a friendly and open group is preferred [9] as users feel more included [10]. Volonte et al. furthermore found that VAs with a positive attitude can be used to foster simulated social dialogues as users are more encouraged to engage in a social interaction [39]. Based on Cafaro et al., the out-group behavior, defined as the non-verbal behavior towards the user as an outstander who is about to join the group, has a great influence and should be modeled with care [9].

Besides the previously mentioned works, different applications are available focusing on user-agent, user-group or agent-group encounters. In early applications, users can join a group simply by approaching it to a certain interpersonal distance [9, 33]. However, no direct feedback is provided by the VAs, that the user has transitioned from an outstander to a group member. In [4, 27, 39], users approach a single VA to engage in a conversation. Here, the users' proxemic as well as gazing behavior are used as indicator for an interaction intent and a greeting from the VA indicates the start of the conversation. Narang et al. apply the Bayesian Theory of Mind in a mobile context (user walks among a crowd of pedestrians) to infer the user's intent of engaging in a face-to-face interaction with a specific VA or to pass-by. Thereby, the users' trajectories, their proxemics, and their gaze behavior are taken into account. If an interaction intent is classified, the respective VA engages in mutual gaze [26]. In the work of Yang et al., an agent-group interaction is in focus. Based on identifying the shared space of a group in an F-Formation, they generate approach trajectories to join the group without interrupting the conversation [40]. The join is finished, after the agent approached to a certain distance.

In conclusion, different research is available focusing on modeling natural group encounters. However, no general classification scheme was yet presented to distinguish between a user joining a social group or passing-by, while triggering adequate, visible reactions on the VAs. Thus, we propose such a scheme in the following section.

3 JOINING OR PASSING-BY

In this work, we focus on detecting a user's intent on joining or passing-by social groups. To avoid interaction effects, the behavioral design of the social groups is thereby limited to a static context as detailed in Sec. 3.1. Based on this configuration, a classification scheme is presented which infers the user's intent (see Sec. 3.2) and triggers a natural reaction of the respective VAs (see Sec. 3.3).

[1] In mobile contexts the complexity of the basic components is increased as more social cues, e.g., the walking speed and exact trajectory, need to be considered.

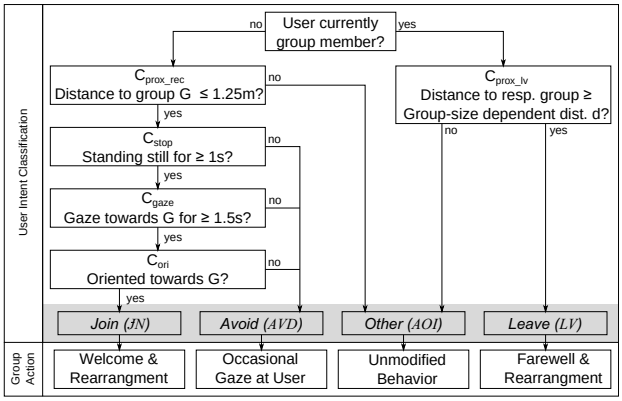

Figure 2: Classification scheme to infer the user's intent and to decide on the individual social groups' next actions.

3.1 Free-Standing, Conversational Groups

To simulate a social gathering in VR, we organize the embedded VAs in free-standing, conversational groups [1] as shown in Figure 1.

We arrange the VAs in a circular shape (see Fig. 4(a)), as common for groups of three and more people [20]. To account for the personal space [16], we introduce interpersonal distances between neighboring VAs based on the outer phase of the personal zone. More precisely, the distances are in a range of .9m to 1.2m, indicating a social bond between the group members. By choosing different values for the individual pairs, the group shapes are asymmetric and thus more natural.

The groups themselves are modeled in a stationary context. Thus, they do not move, split, or merge. To this end, they only increase or decrease in size when the user decides to join or to leave. As the circular arrangements are closed orientations [33], both user actions require adequate group member reactions. During a join, thus a transition from being an outstander to being a group member, the group is required to actively open space for the user. After the user leaves, the group members will automatically close up the free space.

Although remaining in place, the VAs are quasi-dynamic [40] to express liveliness. Therefore, a set of idle animations is used as well as greeting and farewell gestures, and changes in body orientations. Furthermore, the VAs are engaged in scripted conversations. Based on works of Jan et al. [18, 19], the role of the speaker changes as in natural conversations, accompanied by a natural gaze model.

3.2 Inferring the User's Intent

We consider a set of four user intents $I = \{JN, LV, AVD, AOI\}$. *JN* thereby denotes the user's intent of joining a social group, *LV* represents the intent of leaving a previously joined group, while *AVD* denotes the intent of avoiding a group, i.e., passing by. Finally, *AOI* represents any other intent unrelated to our work's focus.

In order to detect all four intents, we developed a two-stage classification scheme, depicted in Figure 2. The first part is used to analyze the current constellation between a user and the social groups to infer the user's intent. Consecutively, we derive an appropriate behavior for the social groups in the second part (see Sec. 3.3).

Based on theoretical reports (e.g., [6]), different (social) cues can be used to infer the user's intent to interact with a previously unconsidered VA or social group. Thus, our classification is based on four criteria related to different significant resources in understanding human intentions. For intents *AOI* and *AVD* using a single criteria turned out to be sufficient. For the remaining intents, a combination of all four is required, as individual criteria can only be used as indicators.

Proxemics. In social interactions, social space arrangements and thus a user's proxemic behavior towards the embedded VAs is important when inferring the user's intent. Based on the literature, we derived two distance-based criteria.

C_{prox_rec}, given in Formula 1, is used to evaluate whether the user is close enough to be consciously recognized by a social group. Thus, it is an indicator for *JN* and *AVD* in contrast to *AOI*, where no group reaction is required. As the user is an outstander at that time, the distance analyzed is the Euclidean distance between the user and the closest VA. Testing a limited range of potential distances based on the following findings, resulted in a suitable threshold of 1.25m:

- Parisi et al. [30] found a minimum pass by distance of approximately .75m in cases of one non-moving pedestrian. Taking into account the sizes of our social groups (cp. findings of [5, 21]), we assume user's will have larger pass-by distances. However, as users also try to minimize the energetic cost of avoidance movements [8], we assume that the distances for a pass-by and a start-to-join are quite comparable.
- Considering the F-Formation [12, 23] (see Fig. 3(a)), two spaces need to be taken into account: the r-space, an area directly outside the group formation, to which outstanders will approach if they intent to join the group [23], as well as the neighboring c-space, an area in which group members start to actively perceive and potentially react to outstanders [12]. However, to the best of our knowledge no common distances for this areas were found, as they depend, i.a., on the personal space [16].
- According to Hall, most social interactions take place in the transition between the 'personal zone' (.45 – 1.20m) and the 'social zone' (1.20 – 3.60m) [16]. The outer phase of the personal distance is thereby used to converse with friends [16]. Thus, we applied it for the in-group proxemics as stated in Sec. 3.1. Additionally, highly organized interactions such as waiting in line [29] take place here, which also accounts for our static social group constellations. Thus, distances within this zone may indicate a *JN*, while we expect slightly larger distances for *AVD*.

$$C_{prox\_rec} = \begin{cases} yes & \text{distance to group} \leq 1.25\text{m} \\ no & \text{otherwise} \end{cases} \quad (1)$$

After the user joined a social group, C_{prox_lv} is used to distinguish between the intents *LV* and *AOI* by evaluating the leaving distance of the user. As the user is a member of the group at that time, the distance taken into account is the Euclidean distance between the user and the group's center The center is located in the middle of the o-space (cp. Fig. 3(a)), so the transactional space to which all group members focus their attention [40]. Due to the group's circular arrangement, the circle's radius will increase with more group members maintaining their personal space. Thus, the leaving distance is group-size dependent, as stated in Figure 2 by the variable d. Formula 2 provides the mathematical description used for the respective distance analysis. Both thresholds were thereby assessed experimentally: per VA .3m increasing the fix distance value in the outer personal zone of 1.0m showed satisfying classification results. This way, the interpersonal

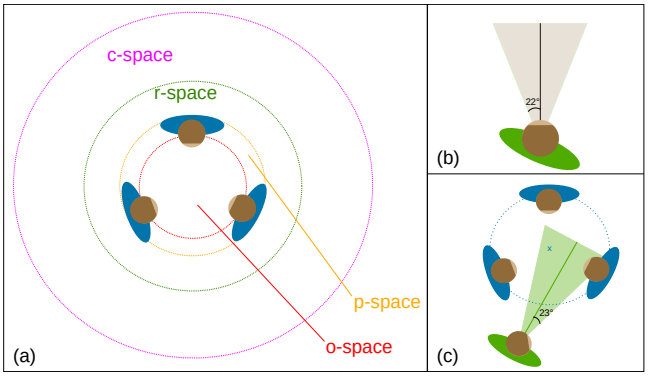

Figure 3: (a) The inner four spaces of the F-Formation (cp. [12]); (b) Approximated eye gaze based on head rotation for C_{gaze}. (c) Torso orientation range used in C_{ori}.

distance towards the two neighboring VAs exceeds the personal and inner social zone. Furthermore, the distance required for a *LV* grows faster compared to the group shape's radius, requiring even larger distances for larger groups. This accounts for an observed "feeling of belonging to a group" even in larger distances to a larger group. Based on our experiments, we furthermore avoided taking gaze or orientation changes into account for *LV*, as they strongly depend on the exact trajectories and behavior chosen by the user on leaving.

$$C_{prox\_lv} = \begin{cases} yes & \text{distance d} \geq 1.0\text{m} + (n \times .3\text{m}); \\ & \text{with } n = \text{no. of VAs in group} \\ no & \text{otherwise} \end{cases} \quad (2)$$

Stops. In a related context, Cafaro et al. introduce the *stop-distance* for stationary conditions [9]. It describes that, besides approaching potential interactants to a given distance, stopping is required to transition from an approaching phase to a direct interaction. Thus, C_{stop}, given in Formula 3, evaluates whether such a stop is performed: While accounting for jitter, the user needs to stay within a deadzone of .25m for 1s or longer. If so, we have an indicator for intent *JN*. Both thresholds are thereby assessed experimentally.

$$C_{stop} = \begin{cases} yes & \text{standing still for} \geq 1.0\text{s} \\ no & \text{otherwise} \end{cases} \quad (3)$$

Gazing. Repeated and prolonged gazing towards an object is a strong social cue for a user's intent to interact with the respective object [17]. As indicated by Narang et al., this also holds true for interactions with VAs [26]. Although introducing instability, using a simplified gaze model focusing only on the user's current gaze [26] instead of the history of gaze [17] has been proven effective. Thus, we use this insight for our gaze criterion C_{gaze}, given by Formula 4. As we expect users to directly gaze towards the objects of interest, instead of subtle gazing from the corners of their eyes, we approximated the eye gaze by the head rotation. To account for the human's horizontal, central field of vision (binocular field) [37], we furthermore introduced an opening angle around the derived viewing direction, shown in Figure 3(b). Experiments indicated an angle of about 22° per eye from the center of the horizontal axis to be suitable. The threshold of 1.5s for the gaze duration, indicating the intent *JN*, was assessed experimentally (cp., [26]). Furthermore, gaze switching between different members of the same social group is allowed, while looking somewhere else leads to a reset of the timer.

Figure 4: On joining, (a) the group's formation *[red circle]* is (b) enlarged *[green circle]* and VAs greet *[right female]*. (c) After following the conversation, (d) leaving triggers a farewell *[females]* and a return to the initial formation *[red circle]*.

$$C_{gaze} = \begin{cases} yes & \text{gaze duration at a social group} \geq 1.5\text{s} \\ no & \text{otherwise} \end{cases} \quad (4)$$

Orientation. With C_{gaze} evaluating only the user's head orientation, the remaining body is neglected. However, the body orientation is a significant component to infer a user's dominant intent: Research showed, that being oriented towards an objects or a person indicates the availability and intent of an interaction [34]. Thus, torso orientation is an indicator for *JN*, resulting in Formula 5. As we consider joining a group instead of approaching an individual VA, C_{ori} evaluates the user's orientation towards the center of a social group. To account for jitter and slight deviations, we furthermore experimentally assessed an acceptable orientation range (cf. Fig. 3(c)). Experiments indicated that an opening angle of 23° is suitable in our scenarios.

$$C_{ori} = \begin{cases} yes & \text{torso oriented towards social group center} \\ no & \text{otherwise} \end{cases} \quad (5)$$

3.3 Triggered Behavior for Social Groups

As depicted in Figure 2, plausible actions for the social groups are derived from the inferred user's intent. They are used as visual backchannels to model user-aware and natural behavior.

In case of *AOI*, there is no cause to react on the user. Thus, the social groups proceed with their current actions, namely the scripted conversation based on interactions described in [18].

In case of *AVD*, the members of the group being passed by are aware of the user's presence in a close surrounding (cp. r-space). Thus, they signal their user-awareness by means of two actions: Random VAs shortly engage in mutual gaze, taking a brief view on the user, before returning to the gaze pattern used during the conversation. In addition, the VA standing closest to the user turns towards him or her as if about to start an interaction. By this, the VA partially opens the closed formation, offering a possibility to join. If the user turns towards the group, a *JN* is assumed. Otherwise, the VAs turn back.

In case of *JN* (see Fig. 4(a-c)), the group needs to actively make space for the user to include him or her. Thus, the two neighboring

VAs start to open space by turning slightly towards the user. By applying territoriality behavior, the VAs are stepping backward to include the user space-wise. Therefore, the social forces F_{circle} and $F_{proximity}$, introduced in [19], are utilized. The later force maintains the interpersonal distances of each member in a range of .9m to 1.2m. Thus, VAs standing too close are pushed backwards, while F_{circle} keeps them within a circular arrangement. In addition, the user is included in the gazing strategy. In parallel, the conversation shortly stops and a greeting gesture is triggered for up to three random VAs, actively welcoming the newly arrived group member.

In case of *LV* (see Fig. 4(d)), the conversation shortly stops and a goodby gesture is triggered for up to three random VAs. Then all VAs readjust their positions and orientations to those prior to the join. At the same time, the user is excluded from the gazing strategy.

4 PILOT STUDY

While we already did smaller testings to experimentally assess the thresholds required for the classification criteria individually, we conducted a pilot study to evaluate the complete scheme.

4.1 Hypotheses

We expected the following hypotheses to be fulfilled:

- **H1:** *Subjects prefer user-aware VAs when joining a social group.* Users interacting with anthropomorphic VAs expect natural, human-like reactions as backchannel for their presence and actions. Thus, when joining a free-standing, conversational group of VAs, subjects will prefer user-awareness, based on our classification scheme, over VAs who ignore them.
- **H2:** *Subjects prefer user-aware VAs when passing-by a social group.* Based on $H1$, we also expect subjects to prefer user-aware VAs on a close pass-by.
- **H3:** *The subject's intent is inferred correctly.* The tests of our individual classification criteria have been successful. Thus, we expect that our complete scheme has a high success rate in detecting the subject's intent.

4.2 Equipment

The used HTC Vive Pro was tracked at $90Hz$ by means of two tripod-mounted SteamVR Base Stations 2.0 in an area of 4.0m×4.0m ($w \times d$). One Vive controller was attached to the subject's torso via a belt to track the orientation. A second controller was used for navigation.

4.3 Environment & Tasks

The pilot study scene consisted of an open backyard with eight social groups, shown in Figure 5(a). The individual characters were modeled via Character Creator[2] and animations were taken from Adobe Mixamo[3]. To realize the conversations, the Google Text-to-Speech API[4] was used to generate the sound files, which were played through the VIVE's build-in headphones as binaural audio. Additionally, corresponding lip sync was created via iClone[5].

We chose a within-subjects design with the VAs' attention level as independent variable. In *AWARE*, our classification scheme was used to trigger appropriate reactions based on the subject's actions. In contrast, in *IGNORE*, the VAs did not react on the subject's presence.

Figure 5: Impressions of the pilot study: (a) the overall setting also used in the *Explore* task, (b) the instructor, as well as examples of the (c) *Join* and (d) *Pass-by* tasks.

We tested both attention levels successively in three tasks, namely:

Join Task. Subjects were asked to join a specific group at a pre-defined location. To this end, a target point and the final path segment towards the target were embedded as green, semi-transparent highlights, shown in Figure 5(c). Subjects navigated with the controller to the path segment's start, then physically walked the segment till they reached the target point. Afterwards, they left the group again to return to their initial position and the next target location was shown. In total, four joins and leaves had to be performed per attention level.

Pass-by Task. Subjects were asked to pass-by a specific group on a pre-defined path. To this end, a path segment with a target point was embedded as green, semi-transparent highlight, shown in Figure 5(d). Subjects navigated with the controller to the start of the path, then walked the path to it's end. On reaching the target point, the next path segment for another group was shown. In total, four pass-bys had to be performed per attention level.

Explore Task. Subjects were asked to freely explore the scene while joining and passing-by arbitrary groups. Before performing an action, they had to first state their intent as well as the related group(s). To allow an unequivocal group identification, unique group numbers were embedded as text labels, floating above the groups (cp., Fig. 5(a)). The task took 4min per attention level.

4.4 Subjects & Procedure

Three females and thirteen males in the age range of 20 to 30 years (M=22.94, SD=2.38) participated in our pilot study. All sixteen subjects were capable of stereoscopic vision and had normal motor skills. Thirteen of them had already used a VR system before. All subjects were naïve to the purpose of the pilot study.

On entering the lab, subjects were informed about the procedure, gave their informed consent and filled out a demographic questionnaire. After being introduced to the safety regulations, subjects were immersed in the empty study scene. Here, a virtual instructor (see Fig. 5(b)) guided them through the pilot study by explaining all tasks. In the familiarization phase, subjects had to navigate to pre-defined locations in the scene to train real walking with the HMD as well as

[2] https://www.reallusion.com/de/character-creator, last-visited: 2020-09-13
[3] https://www.mixamo.com, last-visited: 2020-09-13
[4] https://cloud.google.com/text-to-speech, last-visited: 2020-09-13
[5] https://www.reallusion.com/de/iclone, last-visited: 2020-09-13

flying via the Vive controller while standing[6]. No body avatar was embedded, however, the correctly positioned virtual representations of the input devices were provided. When feeling comfortable, the pilot study began. The order of the *Join* and *Pass-by* task was randomized, while the summarizing *Explore* task was always conducted last. Per task, the order of the attention level used was also randomized to avoid biases in the results. For the *Join* and *Pass-by* tasks we combined two trials per attention level, allowing the subjects more time to gain impressions of the VAs' behavior, followed by two questions in the immersive environment which had to be answered using the Vive controller. Comparable to the two-interval forced choice (2IFC), subjects had to state which mode (neutral term for attention level, to disguise the pilot study's purpose) they preferred and which they considered more realistic. This had to be done twice per task, resulting in the aforementioned four join and pass-by repetitions. After the *Explore* task, subjects took off the HMD, answered some final questions regarding preferences for the VAs' behavior, and left. In total, the study took about 40min/subject, from which 25 were spent fully immersed.

We logged the subjects' and the VAs' position and orientation continuously throughout the study. Additionally, the inferred user's intent was stored as well as the answers and the action announcements.

4.5 Results

Join & Pass-by task. All 16 trials (4 trials × 2 tasks × 2 attention levels) were identified correctly by our scheme. Subjects had to state their preference after experiencing two successive trials per attention level, so twice per task and level. The results are summarized in Table 1. While there are no significant differences between both attention levels, a slight tendency towards *AWARE* for the *Join* task can be observed. Taking a closer look on the individual answers shows that seven subjects changed some of their answers between the first and second survey in either one or both tasks, resulting in 14 changes (21.9%). Twelve times (85.7%) answers were changed from *IGNORE* to *AWARE*, only twice (14.3%) vice versa. Furthermore, six of those changes (42.9%) were registered for the *Join* task, the remaining eight (57.1%) for the *Pass-by* task.

Table 1: Subjects' preferences in the *Join* and *Pass-by* task.

Question	*Join* Task		*Pass-by* Task	
	AWARE	*IGNORE*	*AWARE*	*IGNORE*
Which behavior do you like most?	65.6%	34.4%	59.4%	40.6%
Which behavior is most realistic?	65.6%	34.4%	59.4%	40.6%

Exploration task. While experiencing attention level *AWARE*, 93 actions (*JN*, *AVD*, and *JN* after *AVD*[7]; *LV* was not counted explicitly but followed each *JN*) were announced and performed. Our positive detection rates are shown in Table 2. In case *JN* and *AVD* was not detected, the intent was mainly classified as *AOI*. By manually checking the criteria for all miss-detected *AVD* we found that eight announced pass-by's have been conducted with an interpersonal distance > 1.25m. While two pass-bys were conducted rather close in a distance of 1.28m and 1.37m, two more were further away (1.88m and 1.98m). The remaining four happened outside the social zone (> 3.6m).

[6]Flying metaphors can also generate realistic virtual locomotion trajectories [28].
[7]Subjects, e.g., walked around a group (*AVD*) to join from the other side (*JN*).

Table 2: Intent detection in the *Explore* task.

Intent	Subjects' Intents		Classification	
	Stated	Detected	Correct	False
JN	53	50	94.3%	5.7%
AVD	32	21	65.6%	34.3%
JN after *AVD*	8	6	75.0%	25.0%
Total	93	77	82.8%	17.2%

(*Detected* indicates correct detections of stated intent.)

Besides, the classification results, questions in the post-questionnaire revealed more insights. For a first set of questions, a 7-point Likert scale was used (1=strongly disagree to 7=strongly agree). Based on the results given in Table 3, subjects felt that the social groups recognized their intent to join (M=6.5, SD=.7) and, less distinct, to pass-by (M=5.9, SD=1.3). After the join, subjects had enough physical space (M=6.5, SD=.7) and felt comfortable with it (M=6.1, SD=.9)

Table 3: Specific questions for attention level *AWARE*. (M denotes the mean, SD the standard deviation, Mdn the median.)

Question	*M*	*SD*	*Mdn*
The groups recognized that I want to join.	6.5	.7	7
The groups recognized that I want to pass by.	5.9	1.3	6
I had enough physical space after I joined a group.	6.5	.7	7
After I joined, I felt comfortable about the distance to others.	6.1	.9	6

In two multiple-choice questions we finally asked subjects to choose those behaviors from a set of possible answers, which they thought have been used in the respective attentional level, listed in Table 4.

Table 4: Rate of subjects' confirmation of certain VAs' behaviors in *AWARE* and *IGNORE* compared to the reality.

The agents ...	Subjects' Selections		Reality Check	
	AWARE	*IGNORE*	*AWARE*	*IGNORE*
... looked at you	81.3%	43.8%	✓	-
... made space for you	100%	-	✓	-
... turned towards you	93.8 %	-	✓	-
... greeted you	75%	-	✓	-
... bid you farewell	62.5%	-	✓	-
... changed their facial expressions	12.5%	18.8%	-	-

5 DISCUSSION & LIMITATIONS & NEXT STEPS

The goal of our pilot study was two-fold: (1) Finding indicators that subjects prefer user-aware VAs during social encounters and (2) proving that our classification scheme works. To this end, our subjects experienced social groups with the attention levels *AWARE* and *IGNORE* while joining them or passing-by in different tasks.

In the ***AWARE*** condition, subjects noticed the VAs' reactions as well as the differences in behavior when joining and passing-by. They stated that the VAs inferred their intents correctly, while mainly referring to the territorial and the turning behavior. Overall, the territorial behavior was liked, as subjects had enough physical space. However, one subject argued, that the process of making space felt like the VAs shied away, which indicates the use of an unnatural and exaggerated animation. Another subject stated, that he had too much space and felt a bit isolated. Based on this, the observation

of a slightly decreased agreement when being asked whether subjects felt comfortable with the interpersonal distances, proves once again, that a user's personality has to be considered when modeling behavioral aspects. Although, e.g., proxemics are correct from an objective standpoint, user's may have different needs and expectations. In contrast to the territorial behavior, the VAs' gestures (welcome, farewell) were not always recognized. On joining the group, the limited field of view in the HMD impede seeing the welcome on neighboring VAs. Turning the back on the group when leaving, had the same effect for the farewell. Furthermore, three subjects (18.8%) missed being looked at, which might be due to the VAs' gaze control as discussed later. Finally, two subjects (12.5%) falsely perceived changes in the facial expressions of the VAs. We assume, that this impression emerged due to the expectation of, e.g., an additional smile as welcome (cp., [10]). Overall, our subjects felt actively noticed by the VAs while recognizing the reactions to their actions.

In the ***IGNORE*** condition, the VAs were supposed to take no notice of the subjects. However, seven subjects (43.8%) reported, that the VAs looked at them. This impression emerged as the VAs' gaze was controlled via head movements missing a careful control of the pupils. While looking back and forth between the other group members, we speculate that especially during a pass-by a gaze towards a group member could have been interpreted as a gaze towards the subject. This shortcoming thus reduced the disparity between both attention levels and needs to be taken into account. Besides the mutual gaze, subjects further reported again the falsely perceived changes in the facial expressions. The other differences between both attentional levels have been recognized correctly. Overall, subjects felt noticed, however largely ignored.

Although both attention levels are thus not perceived as opposed as we intended, we can still gain insight from this pilot study:

Detecting the intent *JN* had a success rate of 100% for the *Join* and of 94% for the *Explore* task. We assume the minor decrease for the latter task was caused by the subjects' approaching trajectories: While the predefined trajectories in *Join* are orthogonal (cp. Fig. 5(c)), subject's may have chosen flatter approaching trajectories, comparable to tangents. As we only observed three misdetections, more evaluations need to be conducted to validate our assumption. Nevertheless, our success rates support **H3** for *JN* and the subsequent *LV*.

Evaluating the **awareness preferences for an inferred *JN*** revealed the following: While subjects behaved passively in the *IGNORE* condition, we observed occasional interjections in the *AWARE* condition. They, e.g., waved back, said hello, or thanked the VAs for making space. This indicates that the *AWARE* condition has a more natural frame, in which users feel comfortable. Furthermore, a slight tendency towards the attention level *AWARE* is given. Although we expected a clearer distinction, we still consider **H1** as supported. This is due to the observation that the preferences shifted towards *AWARE* as more liked condition between both surveys (1st: 56.25%, 2nd: 75%).

Detecting the intent *AVD* had a success rate of 100% for the *Pass-by* and of 65.6% for the *Explore* task. Analyzing the false detections revealed, that eight pass-bys happened in a distance contradicting C_{prox_rec} (> 1.25m): Two pass-bys (1.28m, 1.37m) had been close to our proposed distance threshold. Thus, a larger distance to infer the intent *AVD* is required. Four pass-bys had a distance of > 3.6m, so subjects passed outside the social zone. Here, no reaction by the VAs is required and detecting *AOI* is reasonable. Thus, they can be excluded for the detection rate computation. The two remaining pass-bys happened in the social zone (1.88m, 1.98m). Here, it is questionable whether *AVD* or *AOI* is the more reasonable classification. Thus, more insight on the users' preferences is required. To this end, the *AVD* classification needs to be improved as a detection rate of 71.9% (4 exclusions) or 84.4% (6 exclusions) only weekly supports **H3** for *AVD*.

Evaluating the **awareness preferences for an inferred *AVD*** revealed no difference between *AWARE* and *IGNORE*. We argue that this is due to a perceived user-awareness in both conditions, just differing in strength and the VAs expressing it: In *AWARE* the VA next to the subject turned and engaged in mutual gaze. Additionally, random VAs looked at him or her, which is also perceived in the *IGNORE* condition. Thus, having no preference for either condition might be a hint that subjects expect different reactions on their pass-bys. To this end, more research has to be conducted here as we can neither support nor reject **H2**. First, a truly ignorant behavior needs to be tested in comparison. As some subjects considered the VAs' gazing as creepy and others were irritated by the turnings in *AWARE*, more insight into reasonable reactions to an *AVD* needs to be gained. Here, we expect to see an interaction effect between the interpersonal distance and VAs' reactive behavior: The larger the pass-by distance is, the fewer reactions need to be triggered. Thus, both aspects need to be jointly evaluated in a future study.

Despite promising results in the pilot study, we need to address some **limitations**: The misleading gazing in the *IGNORE* condition and the animation quality have been discussed previously. After improving both factors, we expect a clear preference for the *AWARE* condition. Another shortcoming is our focus on the encounter itself: On joining a group, the users transition from outstanders to groupmembers. However, they remain uninvolved listeners as no direct user-agent-interaction takes place. Although this design was consciously chosen for this first trial, it may influence the user's behavior. Thus, a more natural task like joining a group to ask a question or to engage in the conversation might give more insight. Additionally, our VAs only react on socially-accepted behavior and trajectories. Thus, the system needs to become more robust, in particular in relation to users misbehaving, e.g., walking right into the group's o-space.

After overcoming the aforementioned limitations and evaluating the approach with more subjects from a diverse demographics, there are many avenues for **future work**. First, the static scenario should be evaluated in more detail. Experiments should clarify how the user's perception of the group, e.g., in terms of familiarity, relationship to the entities, or group entitativity [3], impacts the expected groups' reactions to a *Join* or *Pass-by*. In addition, the topic of conversation should be considered, as it might influence, e.g., the gazing dynamics. Finally, evaluating the impact of a body avatar is of interest (cp. [24]). A follow-up step may then extend the scheme to master dynamic scenarios. Considering dynamic groupings, allows testing whether the classification scheme can also be used to infer the intent of VAs walking around. Furthermore, the increased complexity in interactions allows for criteria adaptions in the classification scheme. Considering mobile contexts also motivates further investigations. Due to, e.g., the entitativity or changes in formation, e.g., based on coherent or incoherent group behavior [35], detecting joins or pass-by's will be more challenging. Thus, more criteria need to be added to the classification scheme (cp., [6]), e.g., the user's walking speed and the exact trajectory taken in comparison to the group.

The outcome of the discussion is thus as follows: Although the aforementioned limitations lower the informative value of our approach, we provide two useful **contributions** to the research area of Social VR: First, we present a functional, yet refinable, classification scheme to infer whether a user wants to join and to pass-by a social group, based on the user's proxemics, gaze, and orientation. Building on this, we, secondly, showed that users prefer user-aware and reactive VAs in indirect, non-verbal interactions, which is in line with previous research (e.g., [7, 26, 39]). Thus, we strongly recommend modeling VAs with an appropriate interactive capacity.

6 CONCLUSION

Encounters between users and VAs grouped into free-standing, conversational groups, become more frequent in VR applications. While focusing on a user's decision to join or to pass-by such a group, we worked towards enhancing the interactive capacity of the respective VAs. To this end, we developed a classification scheme inferring a user's intent, to trigger a natural reaction of the group members. A pilot study allows the following conclusion: Our classification based on the user's social cues such as proxemics, gazing and orientation works reliable for joins in a stationary contexts. However, the criteria to detect a pass-by and the consecutive reactions of the VAs need to be refined further. In addition, more dynamic situations should be taken into account in the future: VAs moving around in the stationary context, as well as purely mobile situations, i.e., pedestrian groups.

REFERENCES

[1] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe. 2015. Analyzing Free-Standing Conversational Groups: A Multimodal Approach. In *Proc. 23rd ACM Intern. Conf. on Multimedia*. https://doi.org/10.1145/2733373.2806238

[2] J.N. Bailenson, J. Blascovich, A.C. Beall, and J.M. Loomis. 2003. Interpersonal Distance in Immersive Virtual Environments. *Personal. Soc. Psychol. Bull.* (2003). https://doi.org/10.1177/0146167203029007002

[3] A. Bera, T. Randhavane, E. Kubin, H. Shaik, K. Gray, and D. Manocha. 2018. Data-Driven Modeling of Group Entitativity in Virtual Environments. *Proc. ACM Symp. Virtual Real. Softw. Technol.* (2018). https://doi.org/10.1145/3281505.3281524

[4] A. Best, S. Narang, and D. Manocha. 2020. SPA: Verbal Interactions between Agents and Avatars in Shared Virtual Environments using Propositional Planning. *IEEE Virtual Real. Conf.* (2020). https://doi.org/10.1109/VR46266.2020.00-74

[5] A. Bönsch, S. Radke, H. Overath, L.M. Aschè, J. Wendt, T. Vierjahn, U. Habel, and T.W. Kuhlen. 2018. Social VR: How Personal Space is Affected by Virtual Agents' Emotions. In *IEEE Conf. on Virtual Reality and 3D User Interfaces*. https://doi.org/10.1109/VR.2018.84464800

[6] A. Bönsch, T. Vierjahn, A. Shapiro, and T.W. Kuhlen. 2017. Turning Anonymous Members of a Multiagent System into Individuals. In *IEEE Virtual Humans and Crowds for Immersive Environments*. https://doi.org/10.1109/VHCIE.2017.7935620

[7] A. Bönsch, B. Weyers, J. Wendt, S. Freitag, and T.W. Kuhlen. 2016. Collision Avoidance in the Presence of a Virtual Agent in Small-Scale Virtual Environments. In *IEEE Symp. 3D User Interfaces*. https://doi.org/10.1109/3DUI.2016.7460045

[8] J. Bruneau, A.H. Olivier, and J. Pettré. 2015. Going Through, Going Around: A Study on Individual Avoidance of Groups. *IEEE Comput. Graph. Appl.* (2015). https://doi.org/10.1109/TVCG.2015.2391862

[9] A. Cafaro, B. Ravenet, M. Ochs, H. Vilhjálmsson, and C. Pelachaud. 2016. The Effects of Interpersonal Attitude of a Group of Agents on User's Presence and Proxemics Behavior. *ACM Trans. Interact. Intell. Syst.* (2016). https://doi.org/10.1145/2914796

[10] A. Cafaro, H. Vilhjálmsson, and T. Bickmore. 2016. First Impressions in Human-Agent Virtual Encounters. *ACM Trans. Comput. Interact.* (2016). https://doi.org/10.1145/2940325

[11] M. Chollet, M. Ochs, and C. Pelachaud. 2014. From Non-Verbal Signals Sequence Mining to Bayesian Networks for Interpersonal Attitudes Expression. In *Intern. Conf. on Intelligent Virtual Agents*. Springer. https://doi.org/10.1007/978-3-319-09767-1_15

[12] T.M. Ciolek and A. Kendon. 1980. Environment and the Spatial Arrangement of Conversational Encounters. *Sociol. Inq.* (1980). https://doi.org/10.1111/j.1475-682X.1980.tb00022.x

[13] C. Ennis and C. O'Sullivan. 2012. Perceptually Plausible Formations for Virtual Conversers. *Comp. Anim. Virtual Worlds* (2012). https://doi.org/10.1002/cav.1453

[14] M. Ghinea, D. Frunză, J. Chardonnet, F. Merienne, and A. Kemeny. 2018. Perception of Absolute Distances Within Different Visual. Systems: HMD and CAVE. In *Internat. Conf. Augmented Real., Virtual Real. & Comput. Graph.* https://doi.org/10.1007/978-3-319-95270-3_10

[15] M. Gilbert. 1990. Walking Together: A Paradigmatic Social Phenomenon. *Midwest Stud. Philos.* (1990). https://doi.org/10.1111/j.1475-4975.1990.tb00202.x

[16] E.T. Hall. 1966. *The Hidden Dimension*. Garden City.

[17] C.M. Huang, S. Andrist, A. Sauppé, and B. Mutlu. 2015. Using Gaze Patterns to Predict Task Intent in Collaboration. *Frontiers in Psych.* (2015). https://doi.org/10.3389/fpsyg.2015.01049

[18] D. Jan and D.R. Traum. 2005. Dialog Simulation for Background Characters. In *Proc. Int. Conf. Intell. Virtual Agents*. https://doi.org/10.1007/11550617_6

[19] D. Jan and D.R. Traum. 2007. Dynamic Movement and Positioning of Embodied Agents in Multiparty Conversations. *Proc. Int. Conf. Auton. Agents* 1968 (2007). https://doi.org/10.1145/1329125.1329142

[20] A. Kendon. 1990. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. https://doi.org/10.2307/2075490

[21] E.S. Knowles. 1972. Boundaries Around Social Space:" Dyadic Responses to an Invader". *Environment and Behavior* 4, 4 (1972), 437.

[22] J. Lugrin, M.E. Latoschik, M. Habel, D. Roth, C. Seufert, and S. Grafe. 2016. Breaking Bad Behaviors: A New Tool for Learning Classroom Management Using Virtual Reality. *Frontiers in ICT* (2016). https://doi.org/10.3389/fict.2016.00026

[23] P. Marshall, Y. Rogers, and N. Pantidi. 2011. Using F-formations to Analyse Spatial Patterns of Interaction in Physical Environments. In *Proc. Conf. Comput. Support. Coop. Work*. ACM. https://doi.org/10.1145/1958824.1958893

[24] C. Mousas, A. Koilias, D. Anastasiou, B. Rekabdar, and C.N. Anagnostopoulos. 2019. Effects of Self-Avatar and Gaze on Avoidance Movement Behavior. In *Proc. 26th IEEE Conf. Virtual Real. & 3D User Interfaces*. https://doi.org/10.1109/VR.2019.8798043

[25] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. 2010. The Walking Behaviour of Pedestrian Social Groups and its Impact on Crowd Dynamics. *PloS one* (2010). https://doi.org/10.1371/journal.pone.0010047

[26] S. Narang, A. Best, and D. Manocha. 2019. Inferring User Intent using Bayesian Theory of Mind in Shared Avatar-Agent Virtual Environments. *IEEE Trans. Vis. and Comp. Graphics* (2019). https://doi.org/10.1109/TVCG.2019.2898800

[27] S. Ólafsson, B. Bédi, H.E.E. Helgdóttir, B. Arnbjörnsdóttir, and H. Vilhjálmsson. 2016. Starting a Conversation with Strangers in Virtual Reykjavik: Explicit Announcement of Presence. In *Proc. 3rd European Symp. on Multimodal Comm.* 62–68.

[28] A.H. Olivier, J. Bruneau, G. Cirio, and J. Pettré. 2014. A Virtual Reality Platform to Study Crowd Behaviors. *Transp. Res. Procedia* (2014). https://doi.org/10.1016/j.trpro.2014.09.015

[29] E. Pacchierotti, H. Christensen, and P. Jensfelt. 2006. Evaluation of Passing Distance for Social Robots. In *15th Internat. Symp. Robot and Human Interactive Comm.* IEEE. https://doi.org/10.1109/ROMAN.2006.314436

[30] D.R. Parisi, P.A. Negri, and L. Bruno. 2016. Experimental Characterization of Collision Avoidance in Pedestrian Dynamics. *Physical Review E* (2016). https://doi.org/10.1103/PhysRevE.94.022318

[31] C. Pedica and H. Vilhjálmsson. 2009. Spontaneous Avatar Behaviour for Social Territoriality. *Proc. Int. Conf. Intell. Virtual Agents* (2009). https://doi.org/10.1007/978-3-642-04380-2_38

[32] B. Ravenet, A. Cafaro, B. Biancardi, M. Ochs, and C. Pelachaud. 2015. Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions. In *Proc. Int. Conf. Intell. Virtual Agents*. https://doi.org/10.1007/978-3-319-21996-7_41

[33] M. Rehm, E. André, and M. Nischt. 2005. Let's Come Together — Social Navigation Behaviors of Virtual and Real Humans. In *Procs. Intell. Technol. Interact. Entertain.* https://doi.org/10.1007/11590323_13

[34] J.D. Robinson. 1998. Getting Down to Business: Talk, Gaze, and Body Orientation During Openings of Doctor-Patient Consultations. *Human Communication Research* (1998). https://doi.org/10.1111/j.1468-2958.1998.tb00438.x

[35] F.A. Rojas and H.S. Yang. 2013. Immersive Human-in-the-Loop HMD Evaluation of Dynamic Group Behavior in a Pedestrian Crowd Simulation that Uses Group Agent-Based Steering. *Proc. 12th ACM SIGGRAPH Int. Conf. VR Contin. Its Appl. Ind.* (2013). https://doi.org/10.1145/2534329.2534336

[36] F. Setti, C. Russell, C. Bassetti, and M. Cristani. 2015. F-Formation Detection: Individuating Free-Standing Conversational Groups in Images. *PLoS One* (2015). https://doi.org/10.1371/journal.pone.0123783

[37] A. Torrejon, V. Callaghan, and H. Hagras. 2013. Panoramic Audio and Video: Towards an Immersive Learning Experience. *Proc. 3rd European Immersive Education Summit* (2013), 51–62.

[38] T. Trescak and A. Bogdanovych. 2017. Case-Based Planning for Large Virtual Agent Societies. *Proc. 23nd Conf. Virtual Real. Softw. Technol.* (2017). https://doi.org/10.1145/3139131.3139155

[39] M. Volonte, Y.C. Hsu, K.Y. Liu, J.P. Mazer, S.K. Wong, and S.V. Babu. 2020. Effects of Interacting with a Crowd of Emotional Virtual Humans on Users '. *IEEE Virtual Real. Conf.* (2020). https://doi.org/10.1109/VR46266.2020.00-55

[40] F. Yang and C. Peters. 2019. App-LSTM : Data-driven Generation of Socially Acceptable Trajectories for Approaching Small Groups of Agents. *ACM Proc. 7th Int. Conf. Human-Agent Interact.* (2019). https://doi.org/10.1145/3349537.3351885