Collective Dynamics

Pushing and Non-pushing Forward Motion in Crowds: A Systematic Psychological Observation Method for Rating Individual Behavior in Pedestrian Dynamics

Helena Lügering[1, 2] · **Ezel Üsten**[1, 2] · **Anna Sieben**[1, 3]

[1] Institute for Advanced Simulation 7: Civil Safety Research, Forschungszentrum Jülich, Jülich, Germany,
E-mail: h.luegering@fz-juelich.de, e.uesten@fz-juelich.de, an.sieben@fz-juelich.de
[2] School of Architecture and Civil Engineering, University of Wuppertal, Wuppertal, Germany
[3] Chair of Social Theory and Social Psychology, Ruhr University Bochum, Bochum, Germany

Received: 22 April 2022 / Last revision received: 14 July 2022 / Accepted: 15 July 2022
DOI: 10.17815/CD.2022.138

Abstract Pushing behavior impairs people's sense of well-being in a crowd and represents a significant safety risk. There are nevertheless still a lot of unanswered questions about who behaves how in a crowded situation, and when, where, and why pushing behavior occurs. Beginning from the supposition that a crowd is not thoroughly homogenous and that behavior can change over time, we developed a method to observe and rate forward motion. Based on the guidelines of quantitative content analysis, we came up with four categories: (1) falling behind, (2) just walking, (3) mild pushing, and (4) strong pushing. These categories allow for the classification of the behavior of any person at any time in a video, and thereby the method allows for a comprehensive systematization of individuals' actions alongside temporal crowd dynamics. The application of this method involves videos of moving crowds including trajectories. The initial results show a very good inter-coder reliability between two trained raters with a 90.5% overlap (KALPHA = .79) demonstrating the general suitability of the system to describe forward motion in crowds systematically and quantify it for further analysis. In this way, pushing behavior can be better understood and, prospectively, risks better identified. This article offers a comprehensive presentation of this method of observation.

Keywords Pushing behavior · forward motion · crowd psychology · observation method · content analysis · rating system

1 Introduction

Imagine a crowd of excited fans waiting to enter a concert hall: There is no queuing system and everyone wants to be the first in the hall, for there are no seat reservations either. If you had a bird's eye view to observe this crowd from above, you would likely get the impression that it is just one big throng in which everyone is pushing and shoving. Examining each person individually, however, reveals that the crowd is not actually homogeneous and not everyone is behaving the same. This paper introduces an observation method which focuses on individual behaviors in such crowds and allows for an appraisal of who is pushing at which moment in time to draw a more differentiated picture. The assessment and evaluation of individual behavior is performed by trained observers using videos of crowds and the extracted trajectories.

Crowded situations are common and happen—at least before COVID-19—almost every day. Just think about the jostling at the train station. As ordinary as it may be, the consequences can be very serious. Pushing behavior not only impairs satisfaction during the crowd experience [1], it also poses a safety risk. Different studies show that high motivation, which often involves pushing and shoving, increases density [2, 3], and reports from real-life scenarios indicate that pushing from behind can lead to life-threatening density and pressure resulting in injuries and fatalities [4]. Although there is broad evidence of cooperative behavior in emergencies [4–7], pushing may also occur during evacuations, which further increases the danger. [8] Several simulations of pedestrian crowds have therefore tried to integrate this behavior [9, 10] but without providing a systematic psychological basis.

Aside from the safety issues, pushing and shoving were generally evaluated as inappropriate and unfair in recent studies with a bottleneck set-up [2, 11]. It is quite surprising, though, that the same participants mentioned these behaviors as the most promising strategies for faster access. Whether individuals actually move forward faster by jostling depends, however, on their strength and the density of the crowd. With respect to the crowd as a whole, it has not yet been conclusively determined whether increasing the pressure by pushing changes the flow through the bottleneck. Although it has been suggested that pushing actually decreases the flow—the so called "faster-is-slower" effect [12, 13] — Haghani et al. [3] found no conclusive evidence for this general occurrence in a review of current experimental literature. Their own experiment, however, indicated that at least strong and aggressive pushing prolongs the egress time in a bottleneck situation.

However, not everyone in a crowd pushes to the same extent. In Adrian et al. [2], the percentage of participants engaged in this behavior varied from 29.2 to 78.6%. Reasons for non-pushing were, for example, avoidance of danger or a general aversion to pushing. Additionally, identification with the crowd may influence pushing behavior—high-identification participants tended to push less and to give more help in a mass evacuation scenario [14]. Also, social norms (e.g., triggered by the spatial organization of the crowd) influence whether pushing is an appropriate behavior or not. Queuing, for example, is a social system where norms prevail that are opposed to pushing [11, 15, 16]. These results show very clearly that pushing is a complex behavior influenced by several factors. Apart from this general decision for or against pushing, it is also natural that any human behav-

ior is not static but dynamic and can therefore change over time. This means, of course, that pushing behavior is also dynamic and sometimes people push only to stop in the next moment. Researchers addressing crowd dynamics have nevertheless tended thus far to address pushing as a constant behavior in a homogeneous crowd. Our proposed rating method takes into account these fluctuating dynamics of pushing and non-pushing.

But before examining these complex dynamics, it is essential to understand which behaviors are included when talking about pushing. According to the Cambridge Dictionary "(to) push" means "to move forcefully, especially in order to cause someone or something that is in your way to move, so that you can go through or past them" [17]. Further, it must be distinguished between intentional and unintentional pushing [18]. Unintentional pushing is the physical reaction to a push from behind that results in one person being pushed forward into another person. In intentional pushing, on the other hand, individuals exert energy themselves to build up forward pressure. In recent studies [2, 11], participants mentioned the use of elbows, arms, or shoulders, as well as pushing to the front and pushing to the side as different forms of (intentional) pushing. Additionally, filling gaps is mentioned as a strategy for faster access. It is debatable whether filling gaps is a form of pushing behavior, as it is less aggressive, but it clearly leads to increased density and people moving forward faster. Consequently, for the purpose of our method, we include filling gaps as a form of pushing. This enumeration of possible forms of pushing strongly suggests that simply distinguishing between pushing and non-pushing is too simple to be helpful. Therefore, our method examines two different gradations of pushing, namely, mild and strong. Adapted to this, we also distinguish two gradations of non-pushing: a simple forward movement "with the flow" and a forward movement that is slower than the crowd as a whole and thus "falling back."

The general idea for the observation and rating method is based on quantitative content analysis as used in psychology and the social sciences [19, 20]. With the help of a complete coding system, this method captures the characteristics of a document. The coding system is created before the analysis and contains precise definitions of the characteristic expressions and assigns numbers to them. The details of the coding system, as well as useful examples and explanations for the coding process, are recorded in the codebook. Furthermore, the document is divided into precise units of analysis. The rating is performed by at least two trained raters, and reliability measures serve to ensure their concurrence. While content analysis was initially developed for text documents (such as newspaper articles, diaries), in recent years it has also been adapted for the analysis of images and video material.

Important steps of content analysis for both text and video analysis are [20]: (a) determination of the analysis material, and definition of units of analysis, (b) design of the coding system based on the literature and research questions, (c) tentative application and revision of the coding system, (d) discussion of the validity of the coding system, (e) training of raters, (f) reliability analysis (inter-coder reliability), (g) complete data collection, and (h) statistical evaluation. In this paper, we present our content analytic method for capturing pushing behavior in crowd videos in a step-by-step fashion (with the exception of the last two steps (g and h)—for an analysis of the data at this level has yet to be performed).

2 Method

The method described here uses videos taken of crowds from overhead in confined areas such as in front of bottlenecks. Trained observers pick out individual people one by one and categorize their behavior in every second. To do this, they use a four-level category system that includes pushing and non-pushing behavior. The method is introduced here with a thorough step-by-step explanation, to facilitate its future use by other research groups.

2.1 Determination of the analysis material and definition of units of analysis

Although pushing behavior has been regularly observed in former experiments, an in-depth approach for defining and grading the behavior has not been one of the most prominent objectives in pedestrian dynamics so far. As a result, there is a wealth of video material that can be potentially "recycled" for constituting a base to analyze the behavior (see for example: Pedestrian Dynamics Data Archive [21]). Any video that contains pedestrians in forward motion can be used. The category system can be applied to experiments with very different crowd dynamics (i.e., fast or slow) because this method includes the entire spectrum of pushing and non-pushing behavior. Every participant can be categorized as to the degree and intensity of their behavior, whether pushing is observed or not.

Individual trajectories must be available or first extracted for the video to be evaluated. The detection is done via PeTrack software [22]. PeTrack was mainly developed for automatic extraction of pedestrian trajectories from video recordings that are captured from cameras with a top-down view for measuring the physical properties of crowds (e.g., density). The category system uses these trajectories for individual pedestrians to provide accurate timing (via frame numbers: 1 second is equal to 25 frames) of starting categories, category shifts, ending categories, and their spatial visualizations. PeTrack was upgraded specifically for the current category system; an annotation command and a feature allowing the video to be played in real time were added to the software (Version 0.8.15) in order to have an accurate-timing comment (rating: category 1 to 4) for a specific person and a specific frame. The txt file output shows the rating with the respective frame that is bound to the respective pedestrian.

The rating is executed in specific frames that contain a starting point, an ending point, or a behavior change, for every pedestrian. However, a human observer needs at least one second in order to comprehend the complex behavior (and its potential change in the next second) of an individual and therefore it does not make sense to use the frame units defined in PeTrack. For the category system, a unit of analysis is consequently defined as the behavior of an individual in one second. The frame rate of PeTrack is, however, 25 frames per second. Therefore, it was decided that the median of frame ratings within one second of one participant would be calculated and used as the minimum unit of the rating measure. The process of the rating of pushing behavior is as follows: After the

experiment video selection, the ptc (PeTrack) files were gathered from the IAS-7 database and every pedestrian in the chosen video was annotated according to their behavior. The starting point was considered the first frame (usually frame 0) in which PeTrack detects the selected pedestrian, and the ending point was set as either in the last frame of the video or when (if) the pedestrian reaches the bottleneck. In the latter case, the ending frame was always annotated as "END."

2.2 Design of the coding system on the basis of literature and research questions

As outlined above, pushing is defined as a behavior that can involve using arms, shoulders, or elbows; or simply the upper body, in which one person actively applies force to another person (or people) to overtake, while shifting their direction to the side or back, or force them to move forward more quickly. Pushing usually correlates with speed acceleration. Our approach also includes using gaps as a form of pushing because this is a form of overtaking. We distinguish two gradations of pushing behavior: mild and strong. Accordingly, we also distinguish two gradations of non-pushing forward motion. As a result, a category system with four categories has been created: (1) falling behind, (2) just walking, (3) mild pushing, and (4) strong pushing; as two pushing (3 and 4) and two non-pushing (1 and 2) categories.

Six different parameters were used for rating individuals according to these categories: the position of their arms and hands; the position of their shoulders and heads; their personal space; their interaction with others; speed and acceleration; and attention to the exit. These parameters have different behavioral outputs depending on which category they are in, as can be seen below.

Falling behind (1) is the most passive category in terms of forward motion (Fig. 1). People in this category use their hands and arms less. Their arms are generally crossed or dropped by their sides, apart from cases in which they were chatting with other people and using their hands to gesture (arms and hands position). They show frequent head movements because their attention is scattered; they can hence focus on non-specific things in their environment (shoulder and head position). They mostly have some distance to the group and minimal physical contact. In most cases, they are at the back of the crowd, but, when they are in the front, they may actively increase the distance to the person in front by slowing down (personal space). They might be actively involved in chatting with other participants (interaction with others). They are slow overall—even stopping in some cases or changing their direction to somewhere different than toward the exit—and obstruct the pedestrian flow (speed and acceleration). They are focused on other people or things in the environment or become distracted via cell phones instead of focusing on the exit (attention to the exit).

The second category, just walking (2), is applied to people who are not pushing but also not as passive as the people in the falling behind (1) category; they are basically just going with the flow (Fig. 1). People in this category have similar properties with the former category as they can have crossed and dropped arm positions, but since they

are mostly within the crowd, they can use their arms close to their upper body to protect against possible pushing behaviors and they may hold onto fixed objects or barriers to stabilize themselves (arms and hands position). They move slowly and methodically, and they can form a penguin-like waddling motion (shoulder and head position). While they are mostly maintaining their position relative to the crowd and staying in their line, they can be in close body contact with others around them if they are jammed or shoved but under normal circumstances they have sufficient space around them to avoid body contact, as they do not actively increase or decrease the distance to others under a length of half a meter (personal space). They sometimes chat while they are walking (interaction with others). They are also slow and steady, and they may let others go first (speed and acceleration). They can focus on protection or the environment while they are walking toward the exit (attention to the exit).

Mild pushing (3) is a genuine pushing category but, as the name implies, a less active category than the fourth (Fig. 1). People in this category actively increase the density of the crowd. They may raise their arms to apply force to the back of other persons or extend their elbows and arms, or even stabilize themselves by holding on to barriers to prevent others from overtaking (arms and hands position). They often move fast and methodically; consequently, they can form a "fast" penguin-like waddling motion (shoulder and head position). They have much more body contact, they tend to close gaps, change their lines, and overtake for faster access, but without applying excessive force. They may be disproportionately close to the next person without trying to overtake as a tailgating movement or as "psychological pushing," or the closeness can even occur out of an affiliation motive such as hugging someone they know (personal space). They mostly do not chat with other people (interaction with others). They are fast, and they actively decrease their distance to others (speed and acceleration). Their attention is focused on the exit or possible gaps providing a better route (attention to the exit).

The last category, strong pushing (4) is created due to the need for an advanced pushing category for extreme cases (Fig. 1). People with strong pushing behavior tend to use their elbows and hands more strongly to create gaps, they can use barriers to pull themselves forward, they may collide with other people or even pull other pedestrians backward, as they are actively changing their position (arms and hands position). They can move sideways and use a shoulder as a plow, and in most cases, they lean forward (shoulders and head position). They have the most physical contact, and they may create some space behind them due to their rapid movement (personal space). They might communicate with others to engage in coordinated pushing (interaction with others). They are fast and accelerate rapidly when possible (speed and acceleration). Like the mild pushers in the former category, the strong pushers' attention is focused on the exit or possible gaps that might provide a better route (attention to the exit).

All actions in these categories are fully observable in overhead video analysis. This does not mean, however, that people show every parameter in their respective category as they move forward. A person does not necessarily use their arms close to their upper body as protection in just walking (2) if there is no pushing behavior around. There might be no coordinated pushing for people in the strong pushing (4) category if the strong pusher is alone. Consequently, people can be annotated and put in a category depending on their

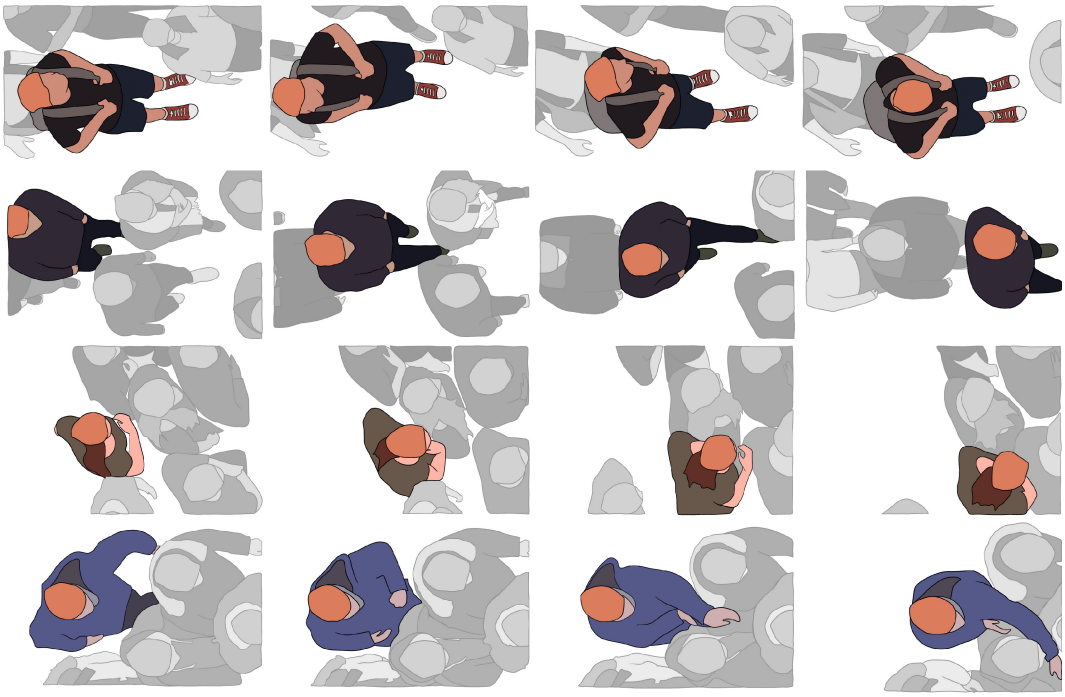

Figure 1 Illustrations of four categories. Each line represents one category. From top to bottom: Category (1) Falling Behind, Category (2) Just Walking, Category (3) Mild Pushing, Category (4) Strong Pushing

prominent behavior even if they do not meet all the parameters.

Another crucial point is that people are not bound to their initial category; as outlined above, they can change their behavior in real life and the category system adapts accordingly to account for these changes. A person might start out as just walking (2) but some time later switch to mild pushing (3) depending on the environment or a shift in motivation. This allows us to describe not only individual differences between people in the crowd but also to capture temporal dynamics.

2.3 Tentative application and revision of the coding system

Once the base structure and the technical properties of the pushing behavior system had been established, raters participated in a series of trials to develop the system further using existing datasets from the project BaSiGo [11, 23, 24] as well as interdisciplinary experiments performed at the University of Wuppertal [2, 25]. All the former experiment video recordings, along with trajectories of each pedestrian, had already been prepared for earlier research and studies and subsequently stored and published in the pedestrian dynamics data archive. The ethic statements for these experiments and recordings can be found in the corresponding papers; no additional ethical approval was necessary for the

current study.

The selected empirical setup for the main trial video was an L-shaped bottleneck scenario, where all participants were instructed to reach the exit with high motivation [11,23]. People were gathered on a platform, each wearing a unique hat (enabling their individual detection from cameras), and were instructed to pass through the bottleneck and exit the platform. Forty pedestrians were randomly selected (out of 123) for the trial dataset and rated accordingly.

The trial ratings revealed that understanding short-term behavior changes is notably challenging: Behavioral shifts of the pedestrians (e.g., category changes from 2 to 3) require more than a second to be comprehended by their actors since there were many examples of momentary behavioral changes for some pedestrians that appear to have happened only by accident (being pushed increases acceleration momentarily in a passive way) or to have been unconscious decisions on the part of the pedestrian (accidental line changing toward a gap), with the former behavior being resumed after one second. It was thus decided that the time gap for a valid and intentional behavioral change should be at least 2 or 3 seconds depending on its context.

2.4 Discussion of the validity of the coding system

Revision of the coding system after some trials revealed some significant points regarding the pushing behavior system. Raters were concerned that they were focused on the observable motivation (having high or low motivation) rather than actual pushing behavior in some cases. While being highly motivated and using strong or mild pushing behavior are potentially highly correlated, the actual behavior can possibly be disregarded while observing the crowd due to the primed motivation of the pedestrian. This vague issue has come up during high-motivation-priming video trials where it was observed that, although most of the pedestrians were highly motivated to reach the bottleneck, not all of them were using pushing behavior. Overall, the main concern was that raters might inadvertently appraise the motivation of the pedestrians instead of their observable pushing behavior.

After careful consideration, raters agreed to conduct the rating process with a context-dependent perspective to avoid this issue. For instance, being fast and accelerated in a calm and slow crowd was agreed to be an indicator for mild (3) or strong pushing (4), but the same behavior can be seen as just walking (2) if the crowd is highly energetic and the average flow speed is similar to the “fast” pedestrian. It is thus helpful to watch the video once before the actual rating to get a feel for the respective context. Raters favored this approach as it is much more accurate for detecting and annotating pushing behavior, as it frames the question to be answered in more concrete terms.

The exactness of timing was also an issue for the consistency between both raters: After several test appraisals, some selected annotations done by two raters were analyzed and found, in fact, to be comparable except for a small time slippage by one or two seconds. It was later decided that the observed behaviors were actually the same but coded differently in time either by mistake or by a time lag caused by the software. Nevertheless, it is only natural for human observers to have minor errors in the timing of their ratings in a highly

detailed and complex dataset, and those minor errors should not be problematic especially if the raters are in agreement about what they have seen. Consequently, raters decided to look more closely at the cases with a time slippage of up to two seconds between ratings and select the proper timing together for the main dataset. This process was called "correction" and was done for all the related cases.

2.5 Training of raters

The same L-shaped bottleneck video [11, 23] was selected for use again as the training dataset for two raters to annotate pedestrians. The remaining participants from the main trial dataset (n = 83, out of 123) were annotated by the raters. The rating was done via PeTrack and a txt output was collected afterward.

The output shows only the respective frames for which a rate comment was inserted (i.e., frame 0 = 3, frame 523 = 4, frame 801 = 3, frame 1792 = END; for participant *number*), hence it always needs to be prepared for data analysis. The first preparation was done manually; the total frame numbers were written in Excel and all the ratings were dragged in between the frames (i.e., frame 0,1,2,3. . . .520,521,522 = 3). After every rating for every frame and every pedestrian was prepared, the median of the ratings for units of seconds was calculated and written accordingly. The final procedure was to assemble all the ratings in one column. These proceedings were done separately for the two annotations of two raters. Later, data columns of two raters were merged (as two columns) and collected in one Excel file. The file was stored for later analyses in IBM SPSS.

2.6 Reliability analysis (inter-coder reliability)

It was decided that the inter-coder reliability should be calculated via Krippendorff's alpha (KALPHA) [26]. Having multiple coders and an ordinal level of measurement (i.e., categories increase from 1 to 4 depending on the behavior), KALPHA was found to be the most effective reliability coefficient for our rating system.

For calculating KALPHA we used a macro by Andrew F. Hayes for IBM SPSS [27]. This macro provides a proper syntax where only the last line must be manually adapted to the respective data set and the required output. This looks as follows: KALPHA judges = judgelist/level = lev/detail = det/boot = z. "Judgelist" contains the names of the raters, "lev" is the measurement level (in our case: ordinal = 2), "det" is a selection of whether there is a need for a more detailed output (0 for only KALPHA value), and "z" is the bootstrapping number (in our case: 10000) [28]. As database for the inter-coder reliability, we used the ratings from the training section. So, N = 83 participants were rated by two independent raters. Please note that N = 43 participants were rated twice by one rater with 4.5 months in between because the first rating was performed before the method had been described in detail for this article. In the process of writing, the categories underwent additional differentiation and clarification, so we decided that both raters should conduct their observations at the same time. As the quality of the first rating thus might remain below what is possible, we repeated it for this paper to demonstrate more accurately the

potential of our system. The second rating round was almost a new one since the rating process is very complex and there was a big time gap between the two ratings. The rater could thus not remember the former ratings and was of course not aware of the rating of the second observer during the process. Finally, the dataset for reliability analysis consisted of 143,172 rated frames. After aggregating 25 frames into one second, 5,717 units of analysis remained. We adjusted 60 units due small time slippages as explained in Sec. 2.4). For this prepared data set, the results show 90.5% overlap between the raters and KALPHA = .79.

Even though De Swert [28] mentioned KALPHA = .80 as an established limit for good reliability, he also stated that lower values (minimum of .67, or even .60 for extreme cases) are acceptable if there are good reasons for it. In our case, there is an extremely large number of analysis units, and our categories further rely on rather minor behaviors which are context dependent and sometimes difficult to detect from above. Additionally, behavioral shifts over time are considered, and the analysis units are somehow dependent from each other (e.g., if one observer sees a shift to mild pushing and therefore changes the rating from 2 to 3 but the other evaluates the behavior differently, the rating does not only differ for one second but immediately for several). Given this complexity of the rating system, a value of .79 is, in our view, more than satisfactory.

Despite this high level of agreement between raters, we nevertheless have partially divergent ratings for some participants. If the data is to be used for further analysis, however, there cannot be two data sets with divergent values. Therefore, the question is how to combine these different values into one value. The calculation of the mean value, for instance, makes no sense for the method (e.g., 2.5 as mean between just walking and mild pushing). Instead, the raters have to reach a later compromise in cases of disagreement. For that purpose, all divergent cases must be observed again and discussed. This leads to a completely consistent data set that can be used to answer the following research questions. It is essential to note that this step may only be performed after the inter-coder reliability has proven to be high enough.

2.7 Preliminary visualization

For visualization of the rating, we took one video from the Pedestrian Dynamics Data Archive [2, 21, 25]. Screenshots are depicted in Fig. 2 and the full video can be found in the ‘Supplementary files’ section. This visualization is only preliminary to illustrate our rating system. More sophisticate forms can be created using special software (e.g., JuPedSim) [29] or including other quantities (e.g., density).

3 Discussion

Pushing behavior impairs people’ sense of well-being in a crowd and also poses a significant safety risk. Nevertheless, to date it has been barely investigated. Following the idea that a crowd is not thoroughly homogenous in behavior and that there can also be changes over time, we developed a rating system of individual behavior in crowds. Prospectively,

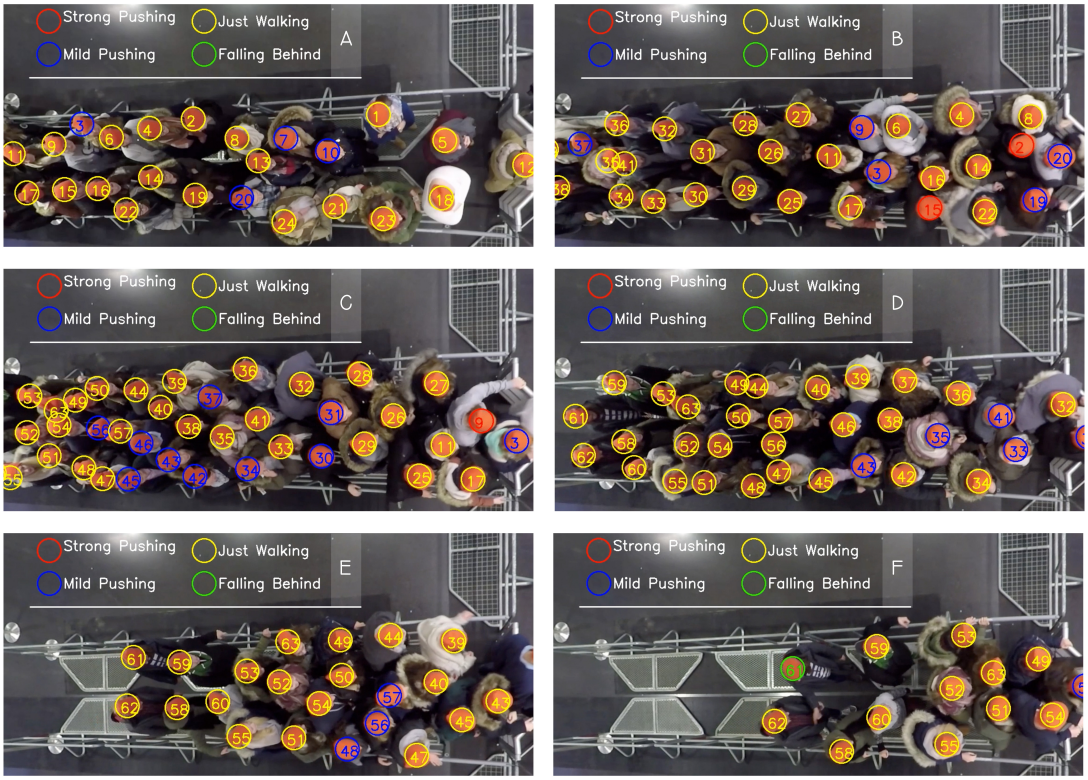

Figure 2 Preliminary visualization of ratings. Screenshots were taken from one exemplary video. Letters (A, B, C, D, E, F) state the order of the crowd flow. Timepoints of the screenshots are: A = 00.00 s, B = 00.08 s, C = 00.16 s, D = 00.24 s, E = 00.32 s, F = 00.40 s.

this can be used to systematize and quantify all kinds of forward motion as we not only capture pushing but also non-pushing behavior. However, since pushing can have various forms, having just a binary distinction would have been too easy. Therefore, we came up with four categories to take this diversity of forward motion into account: (1) falling behind, (2) just walking, (3) mild pushing, and (4) strong pushing. These categories thus enable us to classify the behavior of any person at any time in a video. In this way, we can not only consider the individuality of people but also the temporal dynamics of behavior. Our rating system was built on the scientific basis of content analysis [19,20] and showed a very good inter-coder reliability between two trained raters.

3.1 Limitations

Although the rating system was found to be reliable, it is also worth mentioning its challenges and limitations in order to have a well-rounded perspective on the system. One major concern was noticed during the training process: The rating procedure was too time consuming. Annotating forward motions of numerous pedestrians involves repeated watching of the videos, focusing on a specific person, and determining the exact time periods of behavioral changes. Overall, annotating one pedestrian required at least five

minutes of observation and consideration, as well as inserting the actual rates into the software. Complex cases, however, required as much as ten (or even fifteen) minutes. In order to have a complete annotation of 83 participants, each rater spent at least seven hours preparing the data. Raters spent an additional two hours correcting the data before the statistical analysis could occur (check Sec. 2.4). In the long run, these durations cannot and (more importantly) should not be decreased since the nature of the system depends on detailed observations. Speeding up the rating process might cause human observers to miss valuable information concerning the pedestrians.

The second observed issue was related to the properties of the selected video. Even though the video was high-resolution, image distortion (flattened fish-eye) sometimes made it hard to perceive and determine actual behavior. The software distorts images in this way to depict an accurate trajectory from the pedestrians from the first standing point through the bottleneck, but this also causes pedestrians to be shaped somewhat bizarrely when they move away from the center. The raters tried to adjust their observation and rate accordingly, although some information might have been lost throughout the process due to this situation. In a broader perspective, using only a bird's-eye view could potentially lead to a loss of information, as well, since the observation becomes slightly limited when seen only from this vantage point. Future studies could incorporate secondary cameras with frontal or side angles where it is thought that these could be beneficial.

Finally, the method was limited by the use of only one video for introducing the pushing behavior system. Even though the selected video contains a crowd scenario with varied behaviors, a different kind of environment (i.e., less crowded, high motivation, low motivation) could potentially be constructive for determining the applicability of the system itself. Raters have conducted some informal trials with different videos that suggest that the system is valid in all the cases mentioned. Additionally, investigating multiple exit scenarios or pedestrians moving in different directions could also be beneficial for showing how feasible the system is, although, we firmly believe that the system would be valid in these cases as well. If a crowd scenario contains forward motion of the pedestrians, then the system can potentially be used since it is based on individual observations regardless of the direction of the pedestrian moving. However, crowd contexts such as watching a sport or a music performance cannot be investigated with the current rating system because these situations do not contain forward motion. Nonetheless, regardless of the selected crowd scenario, it has proven beneficial for raters to confer in advance about the category system for each individual experiment and agree on a set of individual examples of the four categories. This minimizes the context effects.

3.2 Practical implications

While on the subject, possible future applications are described below. The first and probably the most prominent future study could be automating behavior detection by utilizing artificial intelligence (AI) [30, 31]. As it was mentioned in the limitations, the rating process is time consuming and laborious, but an automated AI system could dramatically decrease the rating time by assisting raters in appraising clear cases while flagging the ambiguous ones. All in all, the rating system and the actual annotations might be consid-

ered as the beginning of further pushing behavior-related studies since the system opens a door to measure behavior in space and time and can potentially be applied to related research questions. If an automated detection system could be created, later research could use it to acquire the annotations of multiple videos in a short time.

Regarding future research in social and crowd psychology, behavioral effects can be easily observed and measured with the rating system. Observing one person or one group within a crowd is quite difficult due to having a massive amount of information from the environment, but reducing this data to four ordinal categories could be useful for observing what is really happening in the crowd. For instance, behavior propagation can be observed if it exists (i.e., strong pushing behavior propagates between pedestrians over time via exposure) or behavioral clustering can be identified in some specific locations (i.e., mild pushing behavior localizes in front of the bottleneck). The authors are currently working on these research questions in regard to the rating system's future application. These examples could potentially yield crucial insights for crowd management and evacuation studies, as well, since the system allows interested parties to understand pushing and pushing-related behaviors. Ultimately, the rating system should make it easy to recognize if behavior categories affect each other in any way, depending on the time and their position.

Although the rating method is far too time consuming to be directly useful in the application field of crowd management, it directs the focus toward observing individual behavior as a key to understand the strategies people use in crowds. Such knowledge could be very useful for practitioners in the long run since (potentially dangerous) shifts in crowd movement could be better understood. Likewise, using the system can be beneficial in evacuation studies, such as observing the effects of given directions or instructions on the crowd at an individual level. Potentially, researchers can identify unfair or unwanted behaviors and their effects in an evacuation scenario, and then design or model alternate scenarios to avoid dangerous situations. Furthermore, the detailed descriptions of pushing behavior developed for this method could provide a starting point for thinking about automated observation tools for crowds to detect characteristic indicators of problematic behavior.

3.3 Conclusion

Our rating system provides an important and adequate basis for better understanding the complex dynamics of pushing behavior and forward motion in general. In the video we tested, the agreement between two raters was very good, and a consistent and highly reliable dataset can be generated through the subsequent strategy of compromising. In the future, however, the system must prove its suitability for other videos in different contexts (e.g., different motivations, different moving directions or even CCTV footage). An automated solution for speeding up the rating process would be also beneficial. In any case, this idea is worth pursuing since the quantification of pushing behavior is necessary to answer further research questions which will allow researchers to better understand crowds and thus contribute to public safety.

Acknowledgements The authors acknowledge Prof. Dr. Armin Seyfried and Dr. Maik Boltes for providing the PeTrack software, and Deniz Kilic and Tobias Schrödter for updating the software with the necessary functions for category annotation. The authors thank Ahmed Alia for creating video files with ratings/annotations made by authors embedded on pedestrians. The authors thank Panar Ege Usten for creating the category illustrations.

Author Contributions Helena Lügering and Ezel Üsten, contributed equally as the first authors: Development of the rating system, Rating process, Producing and analyzing the data, Writing – Original draft preparation, Revised draft editing / Anna Sieben: General supervision, revision, editing and writing.

References

[1] Filingeri, V., Eason, K., Waterson, P., Haslam, R.: Factors influencing experience in crowds - the participant perspective. Applied Ergonomics **59**, 431–441 (2017). doi:10.1016/j.apergo.2016.09.009

[2] Adrian, J., Seyfried, A., Sieben, A.: Crowds in front of bottlenecks at entrances from the perspective of physics and social psychology. Journal of the Royal Society Interface **17**, 20190871 (2020). doi:10.1098/rsif.2019.0871

[3] Haghani, M., Sarvi, M., Shahhoseini, Z.: When 'push' does not come to 'shove': Revisiting 'faster is slower' in collective egress of human crowds. Transportation Research Part A: Policy and Practice **122**, 51–69 (2019). doi:10.1016/j.tra.2019.02.007

[4] Johnson, N.: Panic at "the who concert stampede": An empirical assessment. Social Problems **34**, 362–373 (1987). doi:10.2307/800813

[5] Cocking, C., Drury, J., Reicher, S.: The psychology of crowd behaviour in emergency evacuations: Results from two interview studies and implications for the fire and rescue services. Irish Journal of Psychology **30**, 59–73 (2009). doi:10.1080/03033910.2009.10446298

[6] Drury, J., Cocking, C., Reicher, S.: Everyone for themselves? a comparative study of crowd solidarity among emergency survivors. British Journal of Social Psychology **48**, 487–506 (2009). doi:10.1348/014466608X357893

[7] Clarke, L.: Panic: Myth or reality? Contexts **1**, 21–26 (2002). doi:10.1525/CTX.2002.1.3.21

[8] Helbing, D., Farkas, I.J., Molnár, P., Vicsek, T.: Simulation of pedestrian crowds in normal and evacuation situations. In: Schreckenberg, M., Sharma, S.D. (eds.) Pedestrian and Evacuation Dynamics, pp. 21–58. Springer (2002)

[9] Henein, C.M., White, T.: Agent-based modelling of forces in crowds. In: International Workshop on Multi-Agent Systems and Agent-Based Simulation, pp. 173–184. Springer (2004)

[10] Kim, S., Guy, S., Hillesland, K., Zafar, B., Gutub, A.A., Manocha, D.: Velocity-based modeling of physical interactions in dense crowds. Vis Comput **31**, 541–555 (2015). doi:10.1007/s00371-014-0946-1

[11] Sieben, A., Schumann, J., Seyfried, A.: Collective phenomena in crowds—where pedestrian dynamics need social psychology. Plos One **12**, 0177328 (2017). doi:10.1371/journal.pone.0177328

[12] Garcimartín, A., Zuriguel, I., Pastor, J., Martín-Gómez, C., Parisi, D.: Experimental evidence of the "faster is slower" effect. Transportation Research Procedia **2**, 760–767 (2014). doi:10.1016/J.TRPRO.2014.09.085

[13] Helbing, D., Farkas, I., Vicsek, T.: Simulating dynamical features of escape panic. Nature **407**, 487–490 (2000). doi:10.1038/35035023

[14] Drury, J., Cocking, C., Reicher, S., Burton, A., Schofield, D., Hardwick, A., Graham, D., Langston, P.: Cooperation versus competition in a mass emergency evacuation: A new laboratory simulation and a new theoretical model. Behavior Research Methods **41**, 957–970 (2009). doi:10.3758/BRM.41.3.957

[15] Mann, L.: Queue culture: The waiting line as a social system. American Journal of Sociology **75**, 340–354 (1969). URL https://www.jstor.org/stable/2775696

[16] Schmitt, B.H., Dubé, L., Leclerc, F.: Intrusions into waiting lines: Does the queue constitute a social system? Journal of Personality and Social Psychology **63**, 806–815 (1992). doi:10.1037/0022-3514.63.5.806

[17] Cambridge dictionary — push (n.d). Retrieved February 1, 2022, from https://dictionary.cambridge.org/de/worterbuch/englisch/push.

[18] Helbing, D., Mukerji, P.: Crowd disasters as systemic failures: analysis of the love parade disaster. EPJ Data Science **1** (2012). doi:10.1140/epjds7

[19] Neuendorf, K.: The content analysis guidebook, 2nd edn. SAGE Publications (2017)

[20] Döring, N., Bortz, J.: Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, 5th edn. Springer (2016)

[21] Institute for Advanced Simulation 7: Civil Safety Research, Forschungszentrum Jülich: Data archive of experimental data from studies about pedestrian dynamics [data archive] (n.d.). URL https://ped.fz-juelich.de/da

[22] Boltes, M., Seyfried, A., Steffen, B., Schadschneider, A.: Automatic extraction of pedestrian trajectories from video recordings. In: Klingsch, W., Rogsch, C., Schadschneider, A., Schreckenberg, M. (eds.) Pedestrian and Evacuation Dynamics 2008, p. 43–54. Springer (2010)

[23] Institute for Advanced Simulation 7: Civil Safety Research, Forschungszentrum Jülich: Entrance 2, entry with guiding barriers (corridor setup) [data set] (2013). doi:10.34735/ped.2013.1

[24] Institute for Advanced Simulation 7: Civil Safety Research, Forschungszentrum Jülich: Entrance 1, entry without guiding barriers (semicircle setup) [data set] (2013). doi:10.34735/ped.2013.2

[25] Institute for Advanced Simulation 7: Civil Safety Research, Forschungszentrum Jülich: Crowds in front of bottlenecks from the perspective of physics and social psychology [data set] (2018). doi:10.34735/ped.2018.1

[26] Krippendorff, K.: Computing krippendorff's alpha-reliability (2011). Retrieved from https://repository.upenn.edu/asc_papers/43.

[27] Hayes, A., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication Methods and Measures **1**, 77–89 (2007). doi:10.1080/19312450709336664

[28] De Swert, K.: Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. Center for Politics and Communication (2012)

[29] Kemloh Wagoum, A., Chraibi, M., Lämmel, G.: Jupedsim: An open framework for simulating and analyzing the dynamics of pedestrians. In: 3rd Conference of the Transportation Research Group of India (2015). URL https://www.researchgate.net/publication/289377829_JuPedSim_an_open_framework_for_simulating_and_analyzing_the_dynamics_of_pedestrians

[30] Alia, A., Maree, M., Chraibi, M.: A hybrid deep learning and visualization framework for pushing behavior detection in pedestrian dynamics. Sensors **22**, 4040 (2022). doi:10.3390/s22114040

[31] Alia, A., Maree, M., Haensel, D., Chraibi, M., Lügering, H., Sieben, A., Üsten, E.: Two methods for detecting pushing behavior from videos: A psychological rating system and a deep learning-based crowd behavior analysis. In: Proceedings of the 10th International Conference on Pedestrian and Evacuation Dynamics (PED2021), Paper No. 62 (2021). URL https://drive.google.com/file/d/1OdQIxcxaoCuQJUjQE4AC0NJ_qzC4mVHH/view