POLITECNICO DI TORINO
Repository ISTITUZIONALE

Metrics vs Surveys: An Analysis for Human-Aligned Benchmarking in Social Robot Navigation

Original
Metrics vs Surveys: An Analysis for Human-Aligned Benchmarking in Social Robot Navigation / Trepella, S., Martini, M., Ostuni, A., Pérez-Higueras, N., Caballero, F., Merino, L., Chiaberge, M.. - (2026), pp. 2483-2489. (2026 IEEE 35th International Conference on Robot and Human Interactive Communication (RO-MAN) Kitakyushu (Jap) 24-28 August 2026) [10.1109/ro-man72286.2026.11714815].

Availability:
This version is available at: 11583/3017392 since: 2026-11-02T10:32:01Z

Publisher:
IEEE

Published
DOI:10.1109/ro-man72286.2026.11714815

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

07 November 2026

Metrics vs Surveys: An Analysis for Human-Aligned Benchmarking in Social Robot Navigation

Stefano Trepella[1], Mauro Martini[1], Andrea Ostuni[1],
Noé Pérez-Higueras[2], Fernando Caballero[2], Luis Merino[2], and Marcello Chiaberge[1]

***Abstract*— Social, also called human-aware, navigation is a key challenge for integrating mobile robots into human environments. The evaluation of such systems is complex, as factors such as comfort, safety, and legibility must be considered. Human-centered assessments, typically conducted through surveys, provide reliable insights but are costly, resource-intensive, and difficult to reproduce or compare across systems. Alternatively, numerical social navigation metrics are easy to compute and facilitate comparisons, yet the community lacks consensus on a standard set of metrics.**

This work explores the relationship between numerical metrics and human-centered evaluations to identify potential correlations. If specific quantitative measures align with human perceptions, they could serve as preliminary benchmarking tools, providing a human-aligned assessment when large-scale surveys are not feasible. Our results indicate that while current metrics capture some aspects of robot navigation behavior, important subjective factors remain insufficiently represented, necessitating new metrics.

I. Introduction

Human-aware robot navigation is a key research area for integrating mobile robots into human environments [1], [2]. Beyond the classical challenges of path planning and obstacle avoidance, human-aware navigation must address qualitative aspects of social interaction, such as comfort, predictability, and personal space, which are difficult to capture with mathematical models [3], [4].

Precisely, these qualitative aspects make the evaluation of human-aware navigation complex. In addition to traditional performance metrics (e.g. time, distance), several quantitative metrics for social navigation have been proposed [5], [6]. These metrics are easy to compute in controlled experiments and enable efficient comparisons between approaches. However, adopting a wide set of contrastive measures is often misleading in clearly understanding the quality of an experiment. Furthermore, there is no agreement on a set of metrics that reliably capture social aspects such as comfort or legibility. Such aspects are typically assessed through participant surveys, which, while informative, are costly, time-consuming, and challenging to scale.

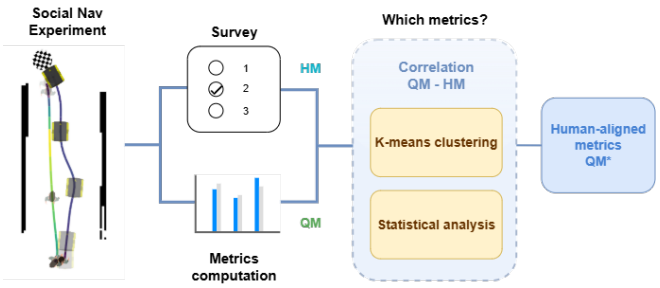


Fig. 1. The proposed framework to infer human metrics (HM) for social navigation experiment using an optimal set of representative quantitative metrics QM*. From the set of experiments, we obtain quantitative metrics from the trajectories and human-provided survey metrics. Then, we use a statistical analysis to determine the correlation between QM and HM, and the optimal set QM*.

This work investigates whether it is possible to infer qualitative aspects from quantitative metrics measured in experiments. We approach the problem by analyzing correlations between quantitative evaluations and qualitative human judgments from surveys in experiments with real robots and human participants. Our goal is to identify quantitative metrics that align with human perceptions, enabling standardized quantitative evaluations that can supplement human-centered validation and provide consistent preliminary benchmarks. A schematic of the proposed framework is illustrated in Fig. 1. The correlation analysis combines K-means clustering with statistical methods to identify sets of quantitative metrics strongly related to human evaluations.

The contributions of this work can therefore be summarized as follows:

1) A novel framework to perform a joint analysis of surveys and quantitative metrics, offering insights on the correlations between them through clustering and statistical tests.
2) The identification of a subset of quantitative metrics that results in the most relevant human-like evaluation trends for social navigation experiments.
3) A dataset of robot and people trajectories in eight different social scenarios collected through real-world experiments with an accurate ground truth and survey evaluation, that can be used for development and bench-

This work was partially supported by the SWIch action (P.R. F.E.S.R.2021/27 - D.G.R. n.19-6962) within the EMPATHY project, and by PoliTO Interdepartmental Centre for Service Robotics (PIC4SeR). It is also partially supported by the Technology Exchange Programme of the EuROBIN project (GA 101070596) funded by the European Union, and the project COBUILD (PID2024-161069OB-C31), funded by the Spanish Ministry of Science, Innovation and Universities, the Spanish Research Agency (MICIU/AEI/10.13039/501100011033) and by ERDF "A way of making Europe".

[1]Department of Electronics and Telecommunications, Politecnico di Torino, 10129, Torino, Italy. stefano.trepella@polito.it, mauro.martini@polito.it, andrea.ostuni@polito.it, marcello.chiaberge@polito.it

[2] School of Engineering, Pablo de Olavide University, Avenida Rectora Rosario Valpuesta, 1. Seville, Spain noeperez@upo.es, fcaballero@upo.es, lmercab@upo.es

marking of future research.

The correlation analysis framework, the dataset, and the survey questionnaire are available in the repository[1].

II. Related Work

In the last decade, several studies have made significant contributions to the definition and evaluation of social navigation in robotics. The studies of Francis et al. [4], or Mavrogiannis et at. [3] make a thorough analysis of the challenges in evaluating social navigation algorithms. Another example is the work of Gao and Huang [5], which provides an in-depth review of methods that quantify a robot's adherence to social norms. Their study examines metrics such as maintaining appropriate interpersonal distances, minimizing discomfort among human co-habitants, and ensuring that paths are socially acceptable. Classical quantitative benchmarks are usually derived from the analysis of the Proxemics of human-human interaction distances [7], as in the review in [8]. Furthermore, other theories of social navigation behavior, based on the famous Social Force Model (SFM) [9], allow quantification of the social impact of robot motion using Social Forces and Social Work, taking into account agents' velocities and directions [10].

In recent years, a new wave of robotics simulators emerged with improved human motion and benchmarks for human-aware navigation: HuNavSim [11], SEAN2.0 [12], SocNavBench [13], Arena 4.0 [14]. These simulators exploit trajectory ground truth and repeatability facilities to compute quantitative metrics for both classical and social navigation.

The most recent research promotes adopting new, comprehensive metrics that integrate efficiency, safety, and human perception into a single score. Among them [6], [15], and [16] propose a suite of novel metrics that aim to capture human-centric perceptions. The latter evaluates robot behavior from the perspective of human comfort and social acceptability, combining objective measurements with qualitative data based on surveys. This dual perspective allows researchers to better analyze whether conventional metrics completely capture critical aspects of human-robot interaction. In addition to defining new metrics, a recent study also highlighted the importance of assigning the correct level of complexity to social scenarios [17].

In general, the most recent studies highlight a remaining challenge: aligning quantitative benchmarks with more subjective human evaluations of robot navigation in social contexts. This work analyzes the correlation between currently used quantitative metrics and survey-based human assessment, posing the question whether these hidden patterns can lead to a new evaluation protocol that integrates human feedback with algorithmic metrics.

III. Data collection

In this section, we describe the steps conducted to collect the social navigation dataset and frame the analysis in a precise evaluation protocol. First, we detail the settings of the experimental activity carried out to record trajectories of robot and people moving in the same environment. Then, the eight social scenarios are defined depicting the human motion, the robot motion, and the maps. Quantitative and human-centric metrics considered for the evaluation are therefore defined, and the statistics and format of the survey activity are finally reported.

[1] https://github.com/PIC4SeR/Social-Nav-Metrics-Matching



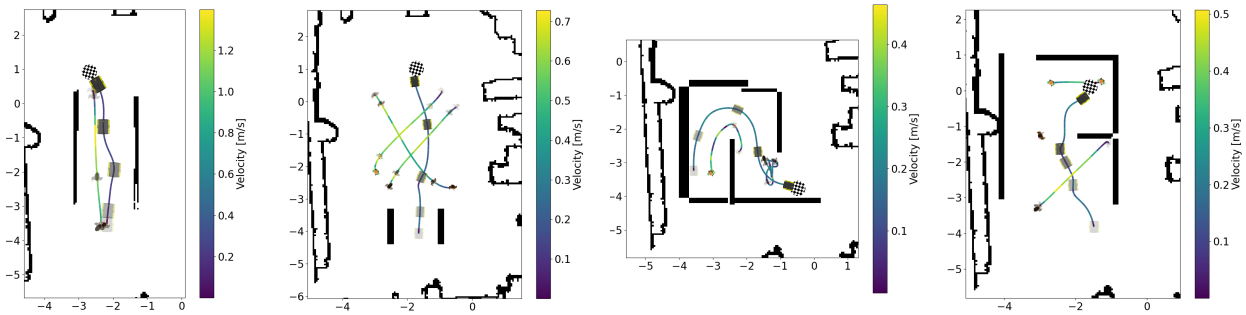
Fig. 2. The laboratory where experiments were performed using a Jackal robot and a VICON tracking system to collect trajectories.

A. Experimental setup

Experimental data have been collected using a real robot in laboratory spaces equipped with a VICON tracking system to record the trajectories of both the robot and humans. The robot adopted is a Jackal from ClearPath Robotics, a typical skid-steer drive robotic platform. An example picture taken in the lab during the experiments is shown in Fig. 2. The effective area used in the laboratory is 4.0×5.0 m, with the first two maps being 2.0×5.0 m, the third being 4.0×3.0 m, and the fourth being 3.0×4.0 m. In the experiments, the robot autonomously navigates using different local planning algorithms, resulting in behaviors with varying levels of social compliance. The controllers have been tuned for each scenario to achieve smoother paths or a less conservative policy, resulting in greater agility and responsiveness. Moreover, in some experiments, the controllers have been coupled with an additional social cost term to exhibit socially compliant trajectories. For instance, a Gaussian costmap around people and the Social Work term defined in Section III-C were used as cost terms in the control function to guide the computation of the robot's velocity commands. The velocity limits of the robot have been set to $v \in [0.0, 0.6]$ m/s and $\omega \in [-1.5, 1.5]$ rad/s for all experiments and the different controllers used.

B. Social Scenarios

Eight different social scenarios have been designed to collect trajectories with robots and humans. The scenarios of *Passing* and *Overtaking* were carried out in Map 1 shown in Fig. 3a. These situations represent two standard interactions in social navigation. Passing involves a robot traversing a corridor while a person walks in the opposite direction, instead overtaking sees the robot trying to surpass a person



(a) Map 1 (b) Map 2 (c) Map 3 (d) Map 4

Fig. 3. Maps used for the experiments and trajectories of robot and humans in four distinct scenarios: Passing, Crossing 3, Narrow turn and Mixed. The velocity scale is indicated by the colorbar on the side, while temporal evolution is expressed through transparency. Other scenarios differ in human motion, as explained in Section III-B.

walking in the same direction. Scenarios *Crossing 1, 2, 3* and *Curious person* were conducted in Map 2 of Fig. 3b. The crossing scenarios involve the robot traveling forward towards its target, while a varying number of people cross its trajectory to reach their destination. Three variations of this case are experimented with: *Crossing 1* has two traversing humans starting from the same side, in *Crossing 2* two people start from opposite sides, and in *Crossing 3* there are three people walking in a cross pattern, two of them being side-by-side.

Maps 3 and 4 were included to test social navigation in more advanced and hybrid scenarios, respectively, as shown in Figs. 3c and 3d. In Scenario *Narrow turn*, the robot must perform a U-shape blind turn with a person appearing from behind the corner and then reach the goal while avoiding a person walking in the same narrow passage. Scenario *Mixed* is a complex scenario obtained by combining different situations. The robot traverses a room with three people, one of them standing still, and the last one walking close to the robot's target. Lastly, the scenario *Curious person* involves a curious person who actively follows the robot and obstructs its path.

In the first seven scenarios, the humans walk at a moderate speed, between 0.4 and 1 m/s, treating the robot as another person and modifying their trajectories only to avoid potential collisions. In contrast, in Scenario *Curious person*, the person chases and disrupts the robot's motion, acting as a curious agent. Three runs of each scenario were performed with different settings of the robot controllers,leading to different social-navigation behaviors.

C. Quantitative metrics

We employ a set of eleven numerical metrics used in the literature and compiled in standard social navigation benchmarks, such as HuNavSim [11]. This QM set is composed of four metrics focused on classic navigation aspects, and seven metrics for social behavior:

- *Time to goal* (TTG) [s]
- *Path length* (PL) [m]
- *Cumulative heading changes* (CHC) [rad]
- *Average robot linear velocity* (ARV) [m/s]
- *Social work* (SW)
- *Social work per second* (SW_s)
- *Average minimum distance to closest person* (AMD) [m]
- *Proxemics occupancy*: intimate space (PR_I), personal space (PR_{PE}), social space (PR_S), public space (PR_{PU}).

CHC represents the total angular distance spanned by the robot during the path. The quantity of social work is derived by the Social Force Model (SFM) of interaction between a crowd of agents proposed in [9]. The social work SW is computed at each time step for the robot according to the following expression, considering the velocity module and direction of each agent around the robot: $SW = W_r + \sum_i W_{p,i}$, where W_r is the sum of the modulus of the robot social force (F_P) and the robot obstacle force (F_O) along the trajectory according to the SFM, while W_p is the sum of the modulus of the social forces generated by the robot for each pedestrian i along the trajectory in a maximum range of 5 meters. Social work is included both as an average per second and as an overall sum throughout the duration of the experiment.

The proxemics metrics are four complementary percentages that represent the time spent by the robot in a given zone: closest to the person (intimate), close to the person (personal), a reasonable distance from the person but still in a range of possible interaction (social), completely away from the agent (public).

D. Human metrics from survey Questionnaire

We have employed a set of four human-centric qualitative metrics (HM) to rank the experiments across relevant aspects, as proposed by Hirose et al. [18]. They are obtained through questionnaires sent to participants using the 5-point Likert scale. These metrics are:

- *Unobtrusiveness* (UO): the robot's ability to move in the environment without being noticed by humans.
- *Friendliness* (FL): the level of social disturbance of the robot in its trajectory.
- *Smoothness* (SM): the robot's smoothness in its trajectory.

- *Avoidance foresight* (AF): the robot's ability to adjust its motion in advance to avoid possible future collisions with moving people.

The human metrics (HM) have been collected through a survey based on real robot experiments conducted in the lab. For each required assessment, the social navigation scenario is described, explaining the robot task, the situation from the pedestrians perspective, and the metrics significance. The collected trajectories of the robot and the agents were converted into GIFs to provide a top-down view of the experiments. These 2D animations are shown in the survey for evaluation, as can be seen in the survey questionnaire form [2]. A total of 24 experiments have been evaluated with the survey. Fig. 3 shows the four maps used for the experiments and trajectories examples obtained in four different runs of Passing, Crossing 3, Narrow turn, and Mixed scenarios. For the special case of Scenario *Curious person*, the unobtrusiveness score has not been collected in the survey, since the intrusive behavior of the human makes it lose relevance. It has been shown in prior work [18] that by using 2D animations, we focus the participant's attention on physical motion signatures such as velocity and relative distances that results to be primary indicators of social intent. The survey included 70 participants from Italy, Spain, and France, aged 20 to 60 years, with a mean age of 31 years and about 40% female participants. The background knowledge in robotics and autonomous navigation has been rated low for almost 50% of people (up to 2 out of 5), and high for about 30% of people (4 or 5 out of 5). The participants independently rated the experiments on a 5-point Likert scale to avoid relative-ranking bias across scenarios.

IV. Analysis

In this section, we describe the methods adopted to process the data extracted from the social navigation experiments and the surveys. We aim to investigate correlations among metrics and propose a new practical evaluation protocol based solely on quantitative aspects.

The framework is summarized in Fig. 1. Metrics correlation with humans scores is investigated to identify the most representative quantitative metrics through a clustering approach, together with a statistical analysis for single-metric correlation.

A. Metrics Pre-Processing

Given the different measurement units and numerical ranges of the adopted QM metrics, the metrics were rescaled between 0 and 1. In this way, it is possible to perform a direct numerical comparison of the metrics across experiments with diverse geometric layouts, numbers of humans, and motion control algorithms. For each scenario, the run with the best value of a particular QM is assigned a value of 1, while the other two are assigned a lower linearly scaled value, depending on their share of that particular metric. The definition of best value varies depending on the specific metric used to evaluate navigation efficiency or safety. For example, Social Work is better when minimized since higher values represent a greater social impact or disturbance to the surrounding pedestrians' intended paths, according to the Social Force Model theory [9]. Similarly, based on Hall's proxemic principles [7], minimizing occupancy in intimate and personal spaces is prioritized to ensure human comfort and to respect social boundaries. Conversely, social space occupancy is considered better when higher, as it reflects the robot maintaining a socially appropriate but safe distance. Similar considerations were applied to all other QM metrics. The HM metrics have also been scaled in $[0, 1]$ by dividing by 5, making these data ready for aggregation, comparison, and clustering.

[2] https://forms.gle/qyKHhzeYnaH8TPvQ7

B. Clustering

The QM and HM metrics data represent two distinct feature sets of the same 24-experiment sample. In this section, we investigate the overall similarity between the two metric sets by grouping the experiments using the QM and HM feature spaces separately. While human labels were available, we specifically chose an unsupervised K-means clustering approach [19] to determine if the inherent, high-dimensional structure of the quantitative data naturally aligns with human categorical perceptions without the bias of supervised training. This enables a robust assessment of whether numerical metrics can autonomously capture the underlying patterns of social interaction that humans observe. The clusters obtained are first analyzed for their internal composition to determine the optimal number of groups, K. Then, clustering accuracy is used to assess whether the QM and HM feature spaces yield the same experimental groups. The internal composition analysis is performed employing the silhouette score, a metric that quantifies the coherence of clusters. The K values tried to group the 24 experimental samples are in the range $[2, 5]$.

The matching alignment of the clustering is evaluated using the Adjusted Rand Index (ARI) [20], which is frequently used in data clustering analysis. ARI is a coefficient that quantifies the similarity between two divisions of a dataset by comparing how data points are assigned to clusters. For our analysis, the labels assigned to the experiments by clustering using the HM survey metrics were treated as the ground truth. The ARI similarity is computed for the entire set of quantitative metrics and for every possible combination of QMs, except the empty set. This process yielded a total of 2.047 combinations. The objective of this analysis is to identify the most prevalent metrics in the combinations that produce the highest ARI. For this purpose, a cumulative ARI score has been computed for each metric, summing the ARI values obtained over the entire QM combination.

The final output, i.e., the cumulative ARI for each metric, represents the relevance of these metrics in achieving the same groupings obtained with the HM features space using K-means clustering.

TABLE I

ANALYSIS OF HM SET

Scenario	Friedman	Wilcoxon signed-rank test
Passing	< 0.001 (***)	< 0.001 (***)
Overtaking	< 0.001 (***)	< 0.001 (***)
Crossing 1	< 0.001 (***)	< 0.001 (***)
Crossing 2	< 0.001 (***)	< 0.001 (***)
Narrow Turn	< 0.001 (***)	0.0040 (**)
Mixed	< 0.001 (***)	< 0.001 (***)
Crossing 3	< 0.001 (***)	< 0.001 (***)
Curious	0.0025 (**)	< 0.001 (***)

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant ($p > 0.05$).

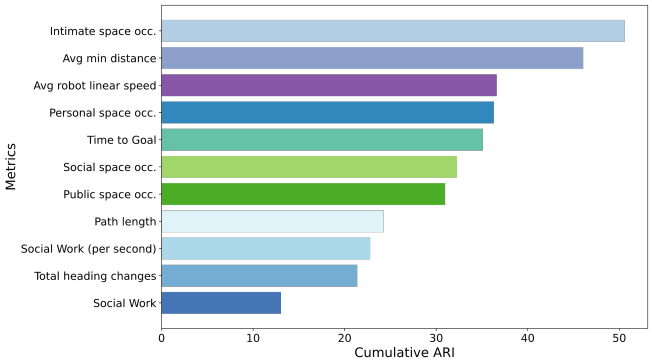


Fig. 4. Histogram of the cumulative ARI. The QM metrics with the highest cumulative ARI are more relevant for clustering experiments, as shown using the HM feature set.

C. Statistical Methods for Evaluating Metrics

The correlation between QM and HM metrics is evaluated using three non-parametric statistical methods: Spearman's Rho, Kendall's Tau, and the Kruskal–Wallis H-test. These methods do not assume normal data distributions and are therefore well suited for ordinal or rank-based evaluations common in social navigation studies.

Spearman: Spearman's Rho evaluates monotonic relationships by converting metric values into ranks and measuring the correlation between paired rank sets.

Kendall: Kendall's Tau measures association by comparing concordant and discordant pairs of ranked observations. The coefficient reflects the proportion of agreement between the orderings of two variables, making it particularly robust for smaller datasets.

Kruskal-Wallis: The Kruskal–Wallis H-test evaluates whether multiple independent groups originate from the same distribution. It compares the rank distributions across groups and determines whether at least one group differs significantly.

V. RESULTS

In this section, results are presented and discussed. First, the correlation between metrics and human-centric evaluations is demonstrated through clustering experiments across both the QM and HM feature spaces. Certain subsets of metrics that maximize the ARI between clusters are therefore identified. Then, individual metrics correlation with the HM set is demonstrated through the statistical analysis. Finally, an overall score comparison demonstrates that quantitative metrics identified through correlation show evaluation trends that align with human assessment.

A. Statistical Validation of Human Ground Truth (HM)

First, we analyze the HM collected. The aim is to mathematically prove that they can clearly identify the different social behaviors across the different scenarios, as expected. We demonstrate that the human perception data collected in the survey is not merely a set of subjective biases, but a consistent and statistically significant measure capable of distinguishing among varying robot behaviors. By validating these responses, we establish the human survey data as the definitive benchmark against which automated quantitative metrics are calibrated. The Friedman test [21] was employed as an omnibus non-parametric test to assess whether evaluation scores differed significantly across the three experimental runs within each scenario. This analysis determines whether different controller configurations produced perceptible differences in navigation quality. Following significant results, post-hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test [22]. In particular, comparisons were performed between Run 3 (expected to exhibit the most socially compliant behavior) and Run 1 (expected to be the least socially compliant). This post-hoc analysis allows identifying which specific behaviors were perceived as significantly different and supports the ranking of navigation strategies in terms of perceived social quality. Table I summarizes the Friedman test results for the aggregated human responses, along with the Wilcoxon results. According to Friedman, the HM survey data maintains a statistically significant score gradient across all eight scenarios ($p < 0.01$), establishing a robust ground truth. Interestingly, in the Narrow Turn scenario, the human participants were statistically slightly less consistent. This suggests that humans find advanced and mixed interactions more complex to evaluate, often expressing a more severe assessment for any trajectory that is only partially social.

B. Clustering comparison

The K-means clustering algorithm is used to study the spatial distribution of experimental samples using QM and HM data as feature spaces. First, we investigate the separability of samples in different numbers of clusters for the HM samples. This internal cluster composition analysis yielded a silhouette score of 0.548 for $K = 2$, while $K = 3$ yielded 0.435, indicating that division into two clusters is more easily distinguishable. Then, we aimed to investigate the matching accuracy between clusters of experiments obtained using two different feature spaces, represented by the HM and QM metrics. This study has been conducted for both $K = 2, 3$, computing a cumulative ARI score as explained in Section IV-B. The results of the clustering accuracy analysis are summarized by the histogram presented in Fig. 4. Regardless of the number of clusters chosen in

TABLE II

THE BEST FOUR QM SETS THAT MAXIMIZE THE ARI SCORES WITH K-MEANS CLUSTERING USING $K = 2$.

Initial ARI	ARI	QM set
-0.04	0.54	ARV, SW_s, PR_I, PR_{PE}, AMD
-0.04	0.42	TTG, ARV, PL, SW_s,PR_I, PR_{PE}
-0.04	0.42	ARV, PR_S, AMD
-0.04	0.31	TTG, ARV, SW_s, PR_{PE}, AMD

K-means, the five metrics with the highest cumulative ARI are: intimate space occupancy, average minimum distance to the closest person, personal space occupancy, average robot linear velocity, and time to goal. This distribution ensures a good balance between the social and trajectory's efficiency metrics. Therefore, both of these aspects contribute to a complete human-like evaluation.

In Table II, the optimal combinations of QM metrics that maximize the ARI are reported, considering $K = 2$. These optimal QM sets significantly increase the ARI compared to using the entire QM set as the feature space. The initial ARI of -0.04 across all 11 quantitative metrics was inadequate to establish a substantial correlation between the two clusters, as it was lower than 0.3. All optimal QM combinations comprise at least three of the most prevalent metrics identified with the cumulative ARI computation. In particular, the average robot velocity and one of the proxemics area occupancy are always present, highlighting the importance of this information for a quantitative evaluation protocol.

C. Statistical Analysis of QM and HM Metrics

A statistical analysis has been conducted to investigate the correlation between the QM and HM sets using a single metric, with the aim of identifying the most representative quantitative metric for each human factor considered. First, we use two non-parametric statistical methods: Spearman's Rho and Kendall's Tau. Both approaches are ideal for data that do not necessarily meet the assumptions of normality, making them well-suited for the ordinal 5-point Likert-scale evaluations used in our survey. We considered correlations that obtain a Spearman strength $|\rho > 0.4|$ with a p-value $p < 0.05$, and a Kendall strength $|\tau > 0.25|$ with a p-value $p < 0.05$. The correlations that satisfy both the strength and the statistical significance criteria defined for the Spearman and Kendall tests are reported on a heat map in Fig. 5 in terms of the average absolute Spearman-Kendall strength.

The QM metrics with the highest correlation with a human criteria (HM) are: Avg-min-distance-person, Proxemics-Intimate-Space, Proxemics-Social-Space, Time-to-goal, and Avg-robot-linear-speed. These results confirm the clustering analysis based on the cumulative ARI score (Fig. 4).

Social Work does not appear among the strongly correlating metrics, as we detected in Fig. 4. Although it is a complex social metric that combines distance and velocity via Social Forces between agents, it does not align with individual HM aspects. Proxemic scores and minimum distance to people correlate with unobtrusiveness, friendliness,

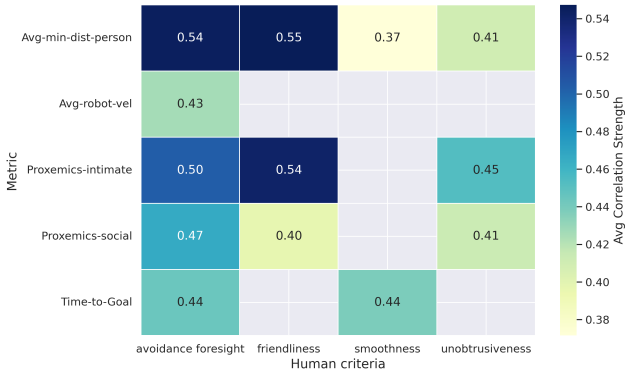


Fig. 5. Heatmap of consistent correlations between single QM and HM metrics, according to both Kendall and Spearman analysis.

and avoidance foresight. Average robot linear velocity relates only to avoidance capability, not to friendliness. Smoothness appears weakly captured by current metrics. While cumulative heading changes were expected to reflect this aspect, only time-to-goal shows sufficient correlation, suggesting it relates to multiple navigation aspects.

We supplemented the correlation study with a group-based statistical analysis using the Kruskal-Wallis H-test. First, while Spearman and Kendall evaluate the monotonic relationship across all 24 experiments, the Kruskal-Wallis test assesses the discriminatory power of the metrics between the three different behavior levels: (Run 1, 2 and 3).

The Kruskal-Wallis Significance Heatmap (Fig. 6) illustrates the -log_{10} p-values, where higher values indicate greater statistical confidence in a metric's ability to distinguish between behavior runs. The results show that the metrics in our optimal QM* set, specifically Proxemics-intimate, Avg-min-dist-person, and Time-to-goal, demonstrate the highest significance across all human criteria. While Smoothness seems to be a subtle aspect poorly captured by metrics, according to both heatmaps, Unobtrusiveness reveal a shift from identifying simple monotonic relationships to establishing high-power statistical discriminators for robot behavior. Notably, the Social Work metric is too noisy what makes it a less reliable metric of social quality. Similarly, the Path length metric is also depreciable. By triangulating these results with our clustering (ARI) and correlation (Spearman/Kendall) data, we provide a robust, human-aligned foundation for the QM* benchmarking protocol.

D. Discussion and limitations

The use of 2D top-down animations representing the environment allowed the survey participants to have a clear understanding of the velocity profiles, relative distances and path smoothness for the displayed scenarios. It should be noted that this methodology lacks the psychological factor of sharing a physical space with the robot.



Fig. 6. Kruskal Wallis significance heatmap ($-log_{10}\ p-values$) illustrating the discriminatory power of quantitative metrics between the 3 experimental runs. Darker cells denote higher statistical confidence in a metric's ability to distinguish between social behavior level. The results confirm the clustering analysis.

VI. Conclusions

This work presents an in-depth analysis of social navigation metrics, investigating existing correlations between current quantitative metrics and qualitative human assessment, with the aim of exploiting hidden data patterns to provide automated benchmarks that align with human judgment. The study proposed a clustering method to identify a QM feature space where experiments are grouped. Then, a statistical analysis of correlation have been performed to inspect individual links between metrics. Our findings show that by selecting relevant QM sets from the correlation results, it is possible to obtain an overall experimental benchmarking that correlates with human-like assessment, aiding in the preliminary screening of social navigation algorithms. Indeed, optimal QM* sets demonstrate clearer scoring trends and behavior identification compared to the complete ensemble of QM. Nonetheless, advanced and mixed social scenarios still represent a challenge due to the complexity of correctly evaluating a navigation experiment that is only partially acceptable, leaving room for future improvement. Although the study identifies strong correlations with human perception, we acknowledge that human surveys remain the gold standard for final validation. The current study is limited by a sample size of 24 real experimental trajectories; therefore, the study proposes a complementary benchmarking tool to human-centered evaluation. Future works will build on these findings to further highlight hidden relationships between QM and HM in larger and mixed datasets. Learning and regression methods will be investigated with the aim of inferring human assessment from quantitative metrics, particularly focusing on the importance and explainability of characteristics.

References

[1] R. Möller, A. Furnari, S. Battiato, A. Härmä, and G. M. Farinella, "A survey on human-aware robot navigation," *Robotics and Autonomous Systems*, vol. 145, p. 103837, 2021.

[2] A. Eirale, M. Martini, and M. Chiaberge, "Human following and guidance by autonomous mobile robots: A comprehensive review," *IEEE Access*, 2025.

[3] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core challenges of social robot navigation: A survey," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[4] A. Francis, C. Pérez-D'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra, H.-T. L. Chiang, M. Everett, S. Ha, J. Hart, J. P. How, H. Karnan, T.-W. E. Lee, L. J. Manso, R. Mirsky, S. Pirk, P. T. Singamaneni, P. Stone, A. V. Taylor, P. Trautman, N. Tsoi, M. Vázquez, X. Xiao, P. Xu, N. Yokoyama, A. Toshev, and R. Martín-Martín, "Principles and guidelines for evaluating social robot navigation algorithms," *J. Hum.-Robot Interact.*, vol. 14, no. 2, Feb. 2025. [Online]. Available: https://doi.org/10.1145/3700599

[5] Y. Gao and C.-M. Huang, "Evaluation of socially-aware robot navigation," *Frontiers in Robotics and AI*, vol. 8, p. 721317, 2022.

[6] J. Wang, W. P. Chan, P. Carreno-Medrano, A. Cosgun, and E. Croft, "Metrics for evaluating social conformity of crowd navigation algorithms," in *2022 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 2022, pp. 1–6.

[7] E. T. Hall, *The Hidden Dimension*. Anchor, Oct. 1990.

[8] S. B. P. Samarakoon, M. V. J. Muthugala, and A. B. P. Jayasekara, "A review on human–robot proxemics," *Electronics*, vol. 11, no. 16, p. 2490, 2022.

[9] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[10] M. Martini, N. Pérez-Higueras, A. Ostuni, M. Chiaberge, F. Caballero, and L. Merino, "Adaptive social force window planner with reinforcement learning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 4816–4822.

[11] N. Pérez-Higueras, R. Otero, F. Caballero, and L. Merino, "Hunavsim: A ros 2 human navigation simulator for benchmarking human-aware robot navigation," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7130–7137, September 2023.

[12] N. Tsoi, A. Xiang, P. Yu, S. S. Sohn, G. Schwartz, S. Ramesh, M. Hussein, A. W. Gupta, M. Kapadia, and M. Vázquez, "Sean 2.0: Formalizing and generating social situations for robot navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 047–11 054, 2022.

[13] A. Biswas, A. Wang, G. Silvera, A. Steinfeld, and H. Admoni, "Socnavbench: A grounded simulation testing framework for evaluating social navigation," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 3, pp. 1–24, 2022.

[14] Volodymyr, L. Kästner, D. Diaz, H. Nguyen, M. Schreff, T. Lenz, J. Kreutz, A. Martban, H. Zeng, and H. Soh, "Arena 4.0: A comprehensive ros2 development and benchmarking platform for human-centric navigation using generative-model-based environment generation," *ArXiv pre-print*, 09 2024.

[15] J. Karwowski and W. Szynkiewicz, "Quantitative metrics for benchmarking human-aware robot navigation," *IEEE Access*, vol. 11, pp. 79 941–79 953, 2023.

[16] P. T. Singamaneni, A. Favier, and R. Alami, "Towards benchmarking human-aware social robot navigation: A new perspective and metrics," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 914–921.

[17] A. Stratton, K. Hauser, and C. Mavrogiannis, "Characterizing the complexity of social robot navigation scenarios," *IEEE Robotics and Automation Letters*, 2024.

[18] N. Hirose, D. Shah, K. Stachowicz, A. Sridhar, and S. Levine, "Selfi: Autonomous self-improvement with reinforcement learning for social navigation," *arXiv preprint arXiv:2403.00991*, 2024.

[19] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[20] D. Steinley, "Properties of the hubert-arable adjusted rand index." *Psychological methods*, vol. 9, no. 3, p. 386, 2004.

[21] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937. [Online]. Available: http://www.jstor.org/stable/2279372

[22] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. [Online]. Available: http://www.jstor.org/stable/3001968