RESEARCH PROPOSAL

Analysis of human behaviour and indoor trajectories from a social perspective

INTRODUCTION

With the increased development of surveillance technologies to identify human behaviour, the necessity to comprehend how individuals interact arises. Group analysis is getting more attention as businesses, and researchers use it to identify people's dynamics in occupied spaces to offer better services, targeting marketing, architectural design and health services (Murino et al., 2017), (Fernandez-nieto et al., 2021). The description of gathering's structures can improve architecture and services such as designing exhibitions in museums, placing interactive screens in retails shops, improving human-robot interactions, and others (Kobayashi et al., 2013; Tröndle et al., 2012; Wineman & Peponis, 2010). Nevertheless, current technologies for detecting groups in closed spaces have sociological, physical, and technological challenges.

Gatherings are mainly studied in sociology to identify leadership (Den, 2017), with increased interest to characterize them in entertainment to take pictures (Gan et al., 2013), in-display design to determine how displays arrangement encourage interaction (Marquardt et al., 2012), and in computer vision to improve robots approaching in conversational groups (Hedayati et al., 2019). However, one of the difficulties in these tasks includes the definition of *physical* and *not physical* areas of interaction. Hall proxemics' theory indicates how humans are spatially predisposed to interact in society (Walters et al., 2009). They are often used to recreate people's behaviour from different distances and activities in which the interaction occurs (Hall, 1966). To investigate the created social structures without direct intervention is necessary to analyse gatherings with more structured representations such as F-Formations. F-Formations are present when "*two or more individuals maintain a spatial and orientational interaction in which the space between them is one with equal, direct and exclusive access*" (A. Kendon, 1976). Sociologists have physically differentiated social interactions in models like F-Formations by direct observation, video analysis, and interviews (Carbon, 2017). The agents and the space within *exclusive access* are not disrupted by others, implying that any social interaction requires communication and *synchronization* to preserve the engagement (Poggi & Errico, 2011). Nevertheless, not much has been said about detecting the *invisible space*, which exists due to social agreements as a *socially occupied space*, product of a synchronized understanding. Situations such as looking at an exhibition in a museum, timetables in a train station, or several people reunited and engaging in the same situations, are examples of how space is *physically occupied* but not disturbed by others due to the appropriation of *invisible space*, created from a social agreement.

Most technological approaches to track individuals focus solely on analysing the spatial aspects of groups by measuring objects and people's location, ignoring the social factors that lead this movement. From a sociological perspective, group interaction has been analysed with field observations, making it difficult to extend it to large scenes (Falk & Dierking, 2016). Other approaches that use tracking technologies include Bluetooth and Wi-Fi to identify location and movement. However, they do not open the analysis to understand the individuals' interaction with the environment or others from their body language (Vom Lehn et al., 2001). The more recent automatic analysis methods use video to detect groups by extracting social cues, focusing mainly on static scenes (Setti et al., 2015). Tracking technologies to describe human behaviour must extract bodily information to identify the conditions for group construction or disarrangement automatically. We hypothesize that with the implementation of more sophisticated tools, such as depth cameras, which can extract body skeleton data (Chen et al., 2018), it is possible to enrich the trajectory information. A trajectory enhancement can be done by integrating social features, formally defined as Social Signals to scale up the sociological analysis in gatherings (Vinciarelli et al., 2009). Social Signals describe the behavioural attitudes

from social intelligence for studying human behaviour to perform group detection and tracking, providing information about "social facts" like social interaction, materialized in body and face direction, location and distance (Vinciarelli & Pentland, 2015). With current approaches implementing technologies to localize objects and people in the *physically occupied space*, and the existing models from sociology to analyse human interaction, *how to automatically detect the socially occupied space from the physical occupied space and derive the social interpretation?*

RELATED WORK

Socially occupied space

Sociology researchers in past years have analysed, defined and structured social encounters from diverse perspectives (Gatica-Perez et al., 2017). Social interaction is defined as an event in which two or more participants establish reciprocal social action or focused interaction, including those with artificial communicative agents (Poggi & Errico, 2011). Interactions can also occur in solitary encounters, in the form of unfocused interactions, in which communication is not preferred and only counts for reciprocal attendance (Adam Kendon, 1990). From Ethnomethodology, social interactions are actions designed to be understandable, self-evident to social members and related to the present context, being aware of others location, actions and presence (Garfinkel, 1991). The social members organize within this context in groups, social units with memberships relations organized by the size, time, formality and degree of belonging (Goffman, 1981). For less socially organized interactions, a particular form of groups, gatherings, is defined as a set of two or more individuals in their immediate presence at a given time, with a situated context-grounded social unit (Bassetti, 2017; Schneider & Goffman, 1964). In both definitions, social units are characterized by the flexibility of the spatial, positional and orientational organization, in which the context of the interaction is the main characteristic going beyond social roles. Researchers have conceptualized gatherings' physical organisation to understand their dynamics, studying interaction with field observations, video, images, and interviews deriving the physical structures of encounters according to their focus (Bitgood, 2006), and distance, so-called proxemics (Hall, 1966). From these characteristics, social units are formally arranged in structures denominated F-Formations (Setti et al., 2015). An F-formation is a space in which "*two or more individuals maintain a spatial and orientational interaction in which the space between them is one with equal, direct and exclusive access"* (A. Kendon, 1976). Thanks to the physical labelling of social encounters, we can *physically* detect the agents and space with *exclusive access,* which are not disrupted by others, implying that any social interaction requires communication and *synchronization* to preserve their engagement (Poggi & Errico, 2011). It is then possible to observe that research has focused mainly on identifying the *physically occupied space* in which the meeting occurs, measurable by observing people and objects. However, not much has been said about the detection of the *invisible space,* which exists due to social agreements as a *socially occupied space* in a synchronized understanding (Poggi & Errico, 2011). With the information provided by the physical environment and, more specifically, the physically social constructs, like F-Formations, we want to provide a solution to interpret the *socially occupied space*. For this task, we need first to identify which visible characteristics are the ones defining social interactions.

Social signals describe a set of behavioural attitudes from social intelligence for studying human behaviour and perform group detection and tracking, directly or indirectly providing information about "social facts" like social interaction, attitudes, relations, and emotions (Vinciarelli et al., 2009). According to the social signal analysis, the sociological features involved in modelling human behaviour are classified in low and high-level social features, including distance, person position, face and body orientation (Poggi & Errico, 2011). We hypothesize that by integrating these physical cues with the implementation of more sophisticated tools, such as the Kinect V2, it is possible to enrich the trajectory data, scaling the sociological analysis for identifying the socially occupied space.

Technological approaches

Different studies have been implemented to evaluate individuals' position in interaction in closed spaces (Kunhoth et al., 2020). For our analysis, these human tracking technologies can be divided into two categories. The first category does not interact with the user directly because its installation is in the surroundings; however, its utility is limited to positional variables and does not directly assess body data to analyse social interactions. For example, Wi-Fi and Bluetooth technologies help to identify device interaction and location. However, they imply large installations and accessing users' devices, giving partial information about people's spatial and body arrangement, in which case, they need additional data such as video recordings or manual records (Marquardt et al., 2012; Yoshimura et al., 2014). LiDAR cameras in museums have similar limitations for assessing enough interaction data with exhibitions, requiring significant processing tasks to get precise trajectories, deriving information mainly about highly concurred areas (Rashed et al., 2016). The second category directly interacts with the user in the form of trackers and markers, interfering with their activities' natural behaviour, especially when users are required to activate beacons to confirm their locations (Dim & Kuflik, 2014).

Moreover, these technologies focus mainly on spatial data, offering only proximity information to identify groups according to their shared space. Trajectories need to be complemented with relevant data to characterize the socially occupied space and offer more context in the sociological analysis of conversational groups such as F-Formations (Vom Lehn et al., 2001). Group interaction has been studied in diverse technologies and fields such as museum visits using forms and manual observation, which implies expensive and lengthy analysis (Bitgood, 2006). Additional techniques imply using cameras to design traditional displays and interactive screens in closed spaces, limiting the analysis of the socially occupied space of the individual or the group concerning others (Marshall et al., 2011), (Kobayashi et al., 2013). Existing computer vision methods use video datasets such as SALSA and Babble to measure orientation to analyse participation in a conversation, ignoring the socially occupied space dynamics that led to these groups' construction (Alameda-Pineda et al., 2015).

Lastly, approaches using the Kinect camera in an egocentric perspective in robots for conversational participation and events limit their analysis to static scenarios and the group's inner characteristics for accounting interactions (Gan et al., 2013; Pathi et al., 2019; Vascon et al., 2016). However, they offer great potential in acquiring relevant social features due to the processed skeleton data, easiness of installation and low costs. We propose using the Kinect V2 camera as a hybrid technology for tracking individuals and collecting body data during trajectories to automate the process of detecting the invisible space, the socially occupied space, in interaction during trajectories.

RESEARCH QUESTIONS AND OBJECTIVES

This research aims to design a technological solution to interpret the socially occupied space automatically. The following research questions are defined:

1. *Which information do we get from the physically occupied space to interpret the socially occupied space?*

   Physically occupied space: it is a taken tangible area whose appropriation is physically represented by a human body or an object.

   Socially occupied space: it is a taken tangible area, in which no physical object represents the appropriation of the physical area, but its occupancy is *recognized* due to a social agreement.

We focus on what information we need to interpret the socially occupied space and how we get it from sensors, in this case, the Kinect V2. We can get the body's location, body orientation, and viewing direction from physically occupied spaces.

2. *How to implement socially occupied space models to analyse their construction?*
   Based on the sensor measurements [RQ1], we transform the sensor measurements into an interpretation of socially occupied space, implementing models like F-Formations.

3. *How do these socially occupied spaces change through time and space? When do they aggregate or disaggregate?* We work on the fusion of several sensors, which enables the observation of a larger space.

The following objectives are defined to answer the previous research questions:

1. Find an approach to technically measure the social features at the spatial level to construct socially occupied spaces.
2. Implement a set of algorithms to process social features and analyse the construction of socially occupied spaces.
3. Evaluate the algorithm comparing the detected socially occupied space with realistic situations in closed spaces.
4. Build a sensor network system to observe movement in a larger space to interpret the socially occupied space across time.

The open challenges identified from the socially occupied space detection process are:

Research Question 1

a. Measure relevant social features.

Research Question 2

a. Identify the shared space for a specific gathering.
b. Identify individuals with their own location and orientation.

Research Question 3

a. Fuse the sensors into one sensor network.
b. Run the Spatio-temporal analysis of socially occupied space.

METHODS AND WORK PACKAGES

We defined three work packages to solve the presented research questions:

1. **WP1: System and technical design**
   Our primary contribution in this package is the information measured from physically occupied spaces with sensors to interpret the socially occupied space. The proposed system implements a depth sensor camera to extract bodily information from individuals in a scene. To identify the low-level social features (distance, gaze direction, aperture, body orientation, radius), the data to be collected include skeleton joints, face, and spatial-temporal information (time-location) to the sensor. The experiments to measure the desired social features, both in artificial and realistic situations, are a) Individual body

orientation and location, b) Dyad body orientation and location, and c) Small gatherings (up to 4 people) body orientation and location. The system's output data is processed to assess location in a global reference system, body orientation, categorization, and face orientation.

2. **WP2: F-Formations and the Socially occupied space**
   The delimitation of the area in which the spatially constructed structures (groups) occur implies processing high-level social features such as effort angle, tightness, symmetry, centroid, and frustum. This package aims to implement a detection algorithm to identify a form of the social model, F-Formations, and describe their spatial components automatically. Current strategies include integrating a joint individual-group base model, which can handle merging and splitting events during trajectories with heatmaps (Gan et al., 2013), delimiting the different n-spaces with proxemics: private, intimate and public spaces. Experiments include data collection from small gatherings in both artificial and realistic situations. The detection algorithm must identify a) Intersection of individuals' field of view, b) Area simultaneously occupied, c) Detecting F-Formations and its categorization.

3. **WP3: Multi-Sensor Fusion**
   Due to limitations involving occlusion and measurement accuracy during the data collection, we proposed to overcome these physical restrictions, the use of multiple cameras in the scene. This procedure implies manipulating large data sets by the algorithm developed in WP1, transforming them into a unique reference system with a complementary trilateration fusion method (Yang et al., 2015) and individual identification to apply the WP2 algorithm. a) Location of an individual before, during and after interaction and its categorization, and b) Aggregation and disaggregation of F-Formations.

EXPECTED RESULTS AND EVALUATION

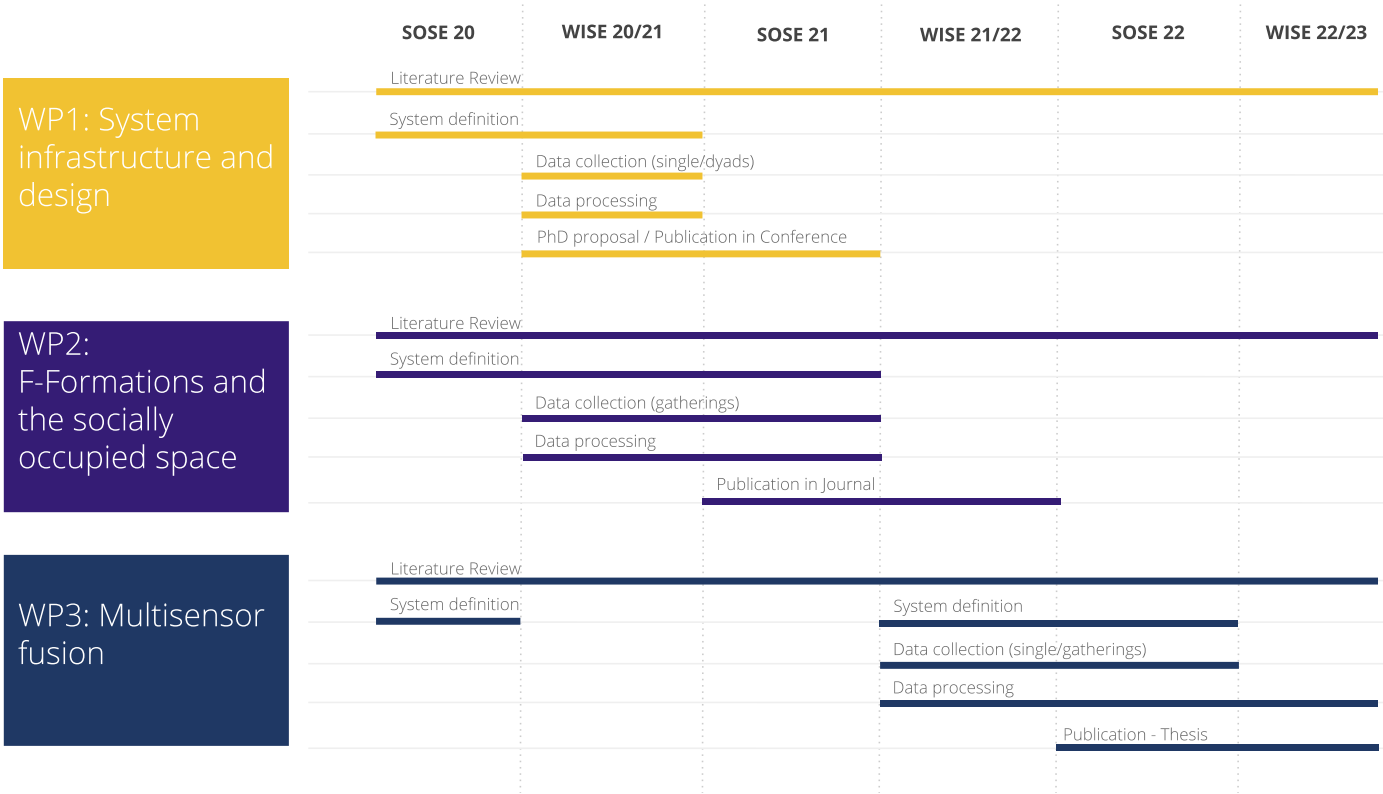
This work's main contribution is the design of a system to interpret the socially occupied space in closed environments automatically. The solution must include:

a. A technical system design and configuration for social features and trajectory data collection for single and multiple depth sensor cameras.
b. A processing algorithm to assess social features and trajectory variables.
c. A detection algorithm to identify spatial configurations for gatherings.
d. A spatial description algorithm to interpret the socially occupied space.

By analysing interactional social models and social spaces' sphere, this work aims to interpret the social interactional space by defining and characterizing the *socially occupied space*. A concept to integrate spatial-temporal information and social features.

Initially, the system's viability to collect relevant social feature variables is evaluated with a precision and recall matrix for the intended categories classification, socially acceptable range, and temporal analysis. Secondly, to analyse the socially occupied spaces and their conceptualization, experts in sociology are invited to review the technological solution in real-life scenarios.

Time Schedule



References

Alameda-Pineda, X., Yan, Y., Ricci, E., Lanz, O., & Sebe, N. (2015). Analyzing free-standing conversational groups: A multimodal approach. *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*, 5–14. https://doi.org/10.1145/2733373.2806238

Bassetti, C. (2017). Social Interaction in Temporary Gatherings: A Sociological Taxonomy of Groups and Crowds for Computer Vision Practitioners. In *Group and Crowd Behavior for Computer Vision* (1st ed.). Elsevier Inc. https://doi.org/10.1016/B978-0-12-809276-7.00003-5

Bitgood, S. (2006). An Analysis of Visitor Circulation: Movement Patterns and the General Value Principle. *Curator: The Museum Journal*, *49*(4), 463–475. https://doi.org/10.1111/j.2151-6952.2006.tb00237.x

Carbon, C.-C. (2017). Art Perception in the Museum- How We Spend Time and Space in Art Exhibitions. *SAGE*. https://doi.org/10.1177/2041669517694184

Chen, N., Chang, Y., Liu, H., & Huang, L. (2018). *Human Pose Recognition Based on Skeleton Fusion from Multiple*. 5228–5232.

Den, Y. (2017). *F-formation and social context: How spatial orientation of participants' bodies is organized in the vast field*.

Dim, E., & Kuflik, T. (2014). Automatic detection of social behavior of museum visitor pairs. *ACM Transactions on Interactive Intelligent Systems*, *4*(4). https://doi.org/10.1145/2662869

Falk, J. H., & Dierking, L. D. (2016). The Museum Experience Revisited. In *The Museum Experience Revisited*. Routledge. https://doi.org/10.4324/9781315417851

Fernandez-nieto, G., Sydney, T., & Martinez-maldonado, R. (2021). *What Can Analytics for Teamwork Proxemics Reveal About Positioning Dynamics In Clinical Simulations ? 5*(April), 1–24.

Gan, T., Wong, Y., Zhang, D., & Kankanhalli, M. S. (2013). Temporal encoded F-formation system for social interaction detection. *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, 937–946. https://doi.org/10.1145/2502081.2502096

Garfinkel, H. (1991). *Studies in Ethnomethodology*. Wiley. https://doi.org/10.1177/0146107916655282

Gatica-Perez, D., Aran, O., & Jayagopi, D. (2017). Analysis of small groups. *Social Signal Processing*, 349–367. https://doi.org/10.1017/9781316676202.025

Goffman, E. (1981). *Forms of Talk*. University of Pennsylvania Press, Incorporated.

Hall, E. T. (1966). *The Hidden Dimension*.

Hedayati, H., Szafir, D., & Andrist, S. (2019). Recognizing F-Formations in the Open World. *ACM/IEEE International Conference on Human-Robot Interaction*, *2019-March*, 558–559. https://doi.org/10.1109/HRI.2019.8673233

Kendon, A. (1976). Spatial organization in social encounters: The F-formation system. *Man Environment Systems*, *6*, 291–296. http://www.researchgate.net/publication/238122903_Spatial_organization_in_social_encounters_The_F-formation_system

Kendon, Adam. (1990). Conducting interaction: Patterns of behavior in focused encounters. In *Conducting interaction: Patterns of behavior in focused encounters.*

Kobayashi, Y., Yuasa, M., & Katagami, D. (2013). *Development of an interactive digital signage based on F-formation system*.

Kunhoth, J., Karkar, A. G., Al-Maadeed, S., & Al-Ali, A. (2020). Indoor positioning and wayfinding systems: a survey. *Human-Centric Computing and Information Sciences*, *10*(1). https://doi.org/10.1186/s13673-020-00222-0

Marquardt, N., Hinckley, K., & Greenberg, S. (2012). Cross-device interaction via micro-mobility and F-formations. *UIST'12 - Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/2380116.2380121

Marshall, P., Rogers, Y., & Pantidi, N. (2011). Using F-formations to analyse spatial patterns of interaction in physical environments. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 445–454. https://doi.org/10.1145/1958824.1958893

Murino, V., Cristani, M., Shah, S., & Savarese, S. (2017). The Group and Crowd Analysis Interdisciplinary Challenge. In *Group and Crowd Behavior for Computer Vision* (1st ed.). Elsevier Inc. https://doi.org/10.1016/B978-0-12-809276-7.00001-1

Pathi, S. K., Kristoffersson, A., Kiselev, A., & Loutfi, A. (2019). F-formations for social interaction in simulation using virtual agents and mobile robotic telepresence systems. *Multimodal Technologies and Interaction*, *3*(4), 1–17. https://doi.org/10.3390/mti3040069

Poggi, I., & Errico, F. D. (2011). *Social Signals : A Psychological Perspective*. https://doi.org/10.1007/978-0-85729-994-9

Rashed, M. G., Suzuki, R., Yonezawa, T., Lam, A., Kobayashi, Y., & Kuno, Y. (2016). Tracking Visitors in a Real Museum for Behavioral Analysis. *Proceedings - 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems, SCIS-ISIS 2016, August*, 80–85. https://doi.org/10.1109/SCIS-ISIS.2016.0030

Schneider, L., & Goffman, E. (1964). Behavior in Public Places: Notes on the Social Organization of Gatherings. *American Sociological Review*, *29*(3), 427. https://doi.org/10.2307/2091496

Setti, F., Russell, C., Bassetti, C., & Cristani, M. (2015). F-formation detection: Individuating free-standing conversational groups in images. *PLoS ONE*, *10*(5), 1–26. https://doi.org/10.1371/journal.pone.0123783

Tröndle, M., Wintzerith, S., Wäspe, R., & Tschacher, W. (2012). A museum for the twenty-first century: The influence of "sociality" on art reception in museum space. *Museum Management and Curatorship*, *27*(5), 461–486. https://doi.org/10.1080/09647775.2012.737615

Vascon, S., Mequanint, E. Z., Cristani, M., Hung, H., Pelillo, M., & Murino, V. (2016). Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, *143*, 11–24. https://doi.org/10.1016/j.cviu.2015.09.012

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*. https://doi.org/10.1016/j.imavis.2008.11.007

Vinciarelli, A., & Pentland, A. S. (2015). New Social Signals in a New Interaction World: The Next Frontier for Social Signal Processing. *IEEE Systems, Man, and Cybernetics Magazine*, *1*(2), 10–17. https://doi.org/10.1109/msmc.2015.2441992

Vom Lehn, D., Heath, C., & Hindmarsh, J. (2001). Exhibiting Interaction: Conduct and Collaboration in Museums and Galleries. *Symbolic Interaction*, *24*(2), 189–216. https://doi.org/10.1525/si.2001.24.2.189

Walters, M. L., Dautenhahn, K., Te Boekhorst, R., Koay, K. L., Syrdal, D. S., & Nehaniv, C. L. (2009). An empirical framework for Human-Robot proxemics. *Adaptive and Emergent Behaviour and Complex Systems - Proceedings of the 23rd Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, AISB 2009*, 144–149.