# UAV-BASED PEOPLE LOCATION TRACKING AND ANALYSIS FOR THE DATA-DRIVEN ASSESSMENT OF SOCIAL ACTIVITIES IN PUBLIC SPACES

JEROEN VAN AMEIJDE[1] and CARSON KA SHUT LEUNG[2]
*[1,2]The Chinese University of Hong Kong*
*[1]jeroen.vanameijde@cuhk.edu.hk, 0000-0002-3635-3305*
*[2]kashutleung@cuhk.edu.hk, 0000-0003-2936-0344*

**Abstract.** In sustainable high-density cities, public spaces play an important role in supporting social and community health and well-being. Amidst ongoing urbanisation, it is of increasing importance to study public space interaction patterns and placemaking processes that contribute to the quality of life of urban residents. This paper reports on the development of a new methodology for the computational tracking and analysis of social activities in urban spaces, using Computer Vision Object Detection (CVOD) techniques to create digitalised pedestrian trajectory data. Referring to concepts from humanistic geography and time geography, our method offers a new platform for data-driven urban place studies, detecting co-presence and social interaction in relation to urban morphology. This paper focuses on the development of Machine Learning protocols, algorithms for tracing and mapping pedestrian trajectories in a georeferenced photogrammetry model, and computational analysis of co-presence. The resulting workflow forms a foundation for future research around detecting, analysing and quantifying behavioural parameters, to evaluate the ability of public spaces to support social interaction and placemaking.

**Keywords.** Public Space Analysis; Pedestrian Location Tracking; Computer Vision Object Detection; Machine Learning; SDG 11.

## 1. Introduction

### 1.1. PUBLIC SPACE, SOCIAL ACTIVITIES AND PLACEMAKING

Amongst the growing awareness of the need to reduce the environmental impact of urbanization, the Compact City concept provides appealing benefits such as enhanced land-use and transportation efficiency, walkability, and other conveniences (Westerink et al., 2013). In high-rise cities such as Hong Kong, the design of public spaces is a crucially important challenge, as they facilitate important everyday activities and socialising, and serve as an extension of people's limited domestic space (Gou et al., 2018). The notion of 'placemaking' emphasizes the role of "context, local conditions, and place-specific culture and experience" in shaping neighbourhood communities and

*POST-CARBON, Proceedings of the 27th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2022*, Volume 1, 293-302. © 2022 and published by the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Hong Kong.

well-being (Williams, 2014, p. 75). Scholars have asserted that place is produced by people's relationships with their environment, geographical behaviour, and the social structures and identities of space and place (Tuan, 1976). 'Secondary interactions' (Jacobs, 1961) can stimulate casual neighbouring, which can lead to social connections, integration, attachment to place (Talen, 1999). Placemaking can contribute to improved social capital, the collective ability to secure resources and opportunities (Friedmann, 2010). Hence, it is of increasing importance to study how sociable public spaces shape interaction patterns and placemaking processes that contribute to the quality of life of urban residents. An increasing number of data-driven studies in recent years has confirmed the association between co-presence and social behaviour (De Stefani and Mondada, 2018; Zakariya et al., 2014). Integrating overlapping insights from the fields of time geography and urban place studies, our research aims to detect behavioural parameters that can be observed, analysed and quantified to evaluate the ability of public spaces to support social interaction and placemaking.

### 1.2. SPATIAL-TEMPORAL PUBLIC SPACE ANALYSIS

Previous studies around documenting activities in public spaces have mixed qualitative and quantitative approaches, following methodologies for the ethnographic study of space outlined by Whyte (1980), Gehl and Svarre (2013) and Low (2016; 2019). These methods combine several observational techniques, including population counts, movement maps and behavioural maps (Low, 2019), and are typically limited in scale across time and geographical space. With the rapid development of information and communication technologies, analysis of increasingly accessible geo-referenced urban data offers a quantitative human-centred approach to capturing the social dynamics of public space. Recent research focuses on the detection of user behaviours and routes based on different types of activity information and Big Data analysis (Biljecki & Ito, 2021; Chen et al., 2016). One category of research, which involves computer vision-based systems to detect and track pedestrians, has been fast developing due to the number of possible applications such as crowd size measurement, transport security, pedestrian traffic management, etc. (Sidla et al., 2006). The research presented in this paper applies a methodology related to this field, but expands it towards the social processes of placemaking in public spaces.

## 2. Methodology

The research presented in this paper employed Unmanned Aerial Vehicle (UAV) mounted cameras, to create a digital capture of an urban space morphology through photogrammetry modelling, and UAV and building-mounted cameras in connection to a newly created pedestrian Real-Time Locating System (RTLS). To obtain RTLS data, multiple cameras were used to obtain fixed position videos of the site from surrounding vantage points. The videos were then analysed with Computer Vision Object Detection (CVOD) techniques to create digitalised pedestrian trajectory data. This data was integrated within the digital models of the urban space using perspective transform algorithms. Quantitative relationships between social activities and public space layout design were then extracted, spatialised and analysed using a customised information processing workflow.

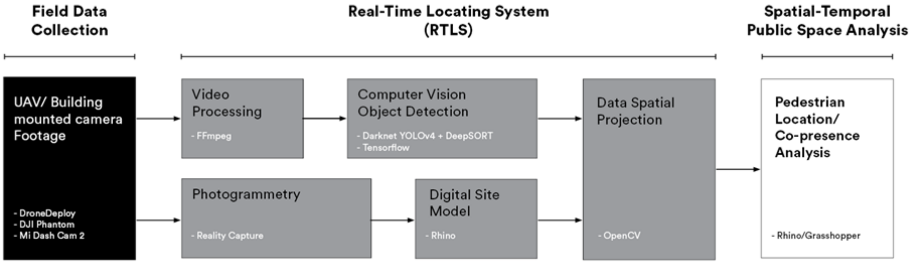
Field Data Collection
Real-Time Locating System (RTLS)
Spatial-Temporal Public Space Analysis
UAV/ Building mounted camera Footage
- DroneDeploy
- DJI Phantom
- Mi Dash Cam 2
Video Processing
- FFmpeg
Computer Vision Object Detection
- Darknet YOLOv4 + DeepSORT
- Tensorflow
Photogrammetry
- Reality Capture
Digital Site Model
- Rhino
Data Spatial Projection
- OpenCV
Pedestrian Location/ Co-presence Analysis
- Rhino/Grasshopper

*Figure 1. Workflow for a RTLS based Spatial Temporal Analysis*

Figure 2 shows the site selected for experimentation, the University Mall at the central campus of The Chinese University of Hong Kong. This area is part of the main access route across the central campus, which connects the main library, administrative buildings, various faculties, and adjacent bus stops.

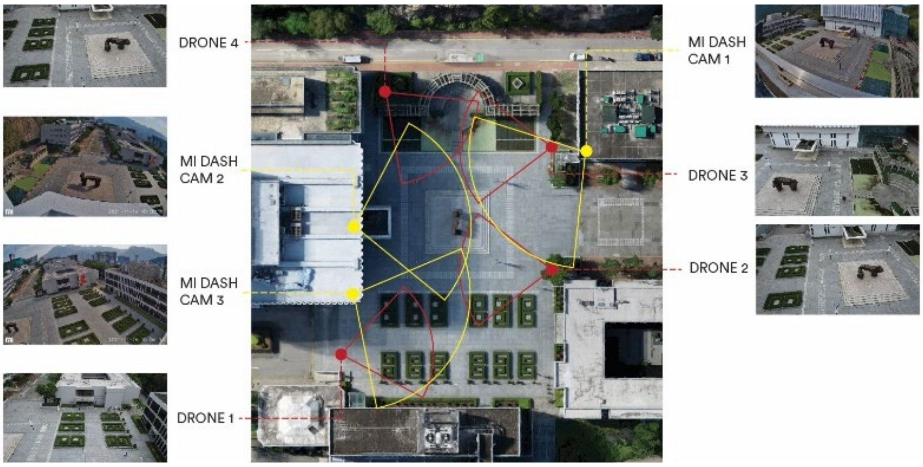

*Figure 2. Multiple Camera System Deployment locations*

## 2.1. PEDESTRIAN DATA COLLECTION

Two methods were developed and tested for obtaining raw bird's-eye view footage of the university mall. The first method used four drones, and the second method used three fixed location cameras (Figure 2). For both methods, the camera angle was calibrated to maximise ground coverage and to exclude the perspective vanishing point. In the first experiment, four DJI Phantom pro 4 v2.0 drones were used to obtain time-synced footages from multiple aerial angles for object tracking purposes to capture pedestrian locations during the evening rush hour. Four drones were dispatched to different holding positions above the square at a flight level of 18m above ground, in which the take-off and landing procedures were designed sequentially to avoid collisions. Within the available battery flight time, the total recording time was 8.5 minutes at 30 frames per second. The per minute record size was 700mb. For the second methodology tests, three Mi Dash Cam 2 car camcorders mounted on tripods and connected to portable battery packs were placed on the roofs of two buildings adjacent to the site. The camcorders were chosen for its ability to loop-record. This type of recording saves one video file per minute, which eases file management and transfer. The Mi Dash Cam 2 has a F1.8, 140° field of view lens which records in a 2K resolution (2560 x 1600). A minute of footage equates to 120mb, therefore with a 128gb microSD card, the camcorder can record up to 17 hours of continuous footage at 30 frames per second. Despite the drones' higher resolution image quality and flexibility in controlling the viewing angle comparing to the building-mounted camcorders, the recording time was severely limited by the battery span. The video stability was also affected by the wind condition and required corrections by the drone operator, compared to the camcorders which remained stable on a tripod and required no human intervention during operation. However, the drone footage yielded a better object tracking result with existing pre-trained object tracking weights, while the footage from the camcorder required a transfer learning approach, training custom object tracking weights for improved accuracy with a one-stage object detection model.

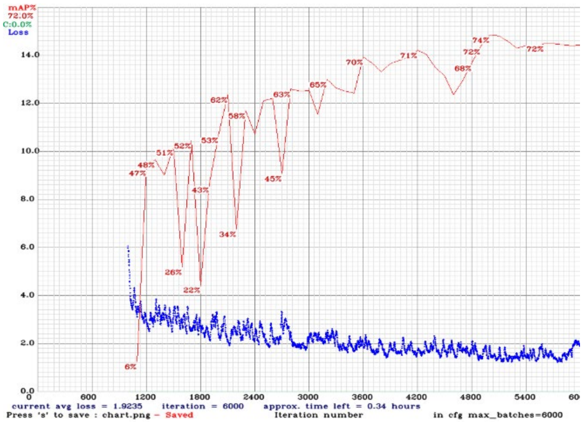

*Figure 3. Training Loss (red line) and mean average precision chart (blue line) chart*

## 2.2. CUSTOM DETECTION WEIGHTS

The type of transfer learning used was parameter transfer under inductive transfer learning, where pretrained parameters between the source domain and target domain model were shared (Pan and Yang, 2009). This training process adds a new detection class to the YOLO (You-Only-Look-Once) v4 pre-trained weights, using selected frames from the collected dataset. This methodology is more time-efficient compared to training a new set of weight from scratch. The Darknet convolution framework, exclusively designed for object detection, was used to process a set of 333 image frames. The images were extracted from the camcorder videos and selectively included different time and angles to maximise the diversity in the learning process. During the detection process, bounding boxes of the target object are drawn on the image using the open-source annotation tool OpenLabeling. This process generates a separate text file, containing the coordinates and its relative detection class number. To train the model, 80% of the data set was used for the training process while 20% was used for evaluation. The return training accuracy rate reached its peak at 76% mean average precision (mAP) after 13 hours of computation time on a Nvidia Geforce RTX 2060 GPU (Figure 3). The dataset used was designed to recognize pedestrians from a bird's-eye viewing angle with a new custom class "human". This avoids the new custom detection class from merging with the pre trained detection class "people" to avoid negative transfer, as the specificities of the dataset it was trained on are unknown.

## 2.3. OBJECT TRACKING

The object tracking process was implemented with YOLOv4, DeepSort, and TensorFlow. YOLOv4 is a one-stage object detection model algorithm that uses convolutional neural networks to perform object detections. Compared to two stage object detection models like R-CNN which have a better detection accuracy but higher inference speed, YOLOv4 was preferred for its lower computational power requirements. YOLOv4 detects the pedestrian positions and draws the relevant prediction boxes per video frame. The detection output of YOLOv4 was then fed into DeepSORT (Simple Online and Realtime Tracking with a Deep Association Metric), to create a continuous object tracker ID (Wojke et al., 2017). It considers information within the bounding box parameters of the detection results of the current and previous frames, to make predictions about the current frame. DeepSORT enable the continuous tracking of the pedestrian even if the detection is lost in certain frames due to a) missed detection or b) the subject passing through covered areas. From the first frame of a successful detection, a unique track ID is assigned to each bounding box which represents the activated detection class. On the top of the box, a confidence value higher is also displayed. Detections which have a confidence value lower than the pre-set threshold will not be displayed. The Hungarian algorithm is used to assign the detections in a new frame to existing tracks. The deep learning process were implemented in TensorFlow, Google's deep-learning software. Comparing to Darknet used in weights training, TensorFlow provides a Python API and is compatible with the Python language, to which a perspective transformation process can be added to compute the real-world location of the tracking point.

## 2.4. TRACKING ACCURACY EVALUATION

To compare the tracking accuracy of the pre-trained YOLOv4 and the self-trained detection weights, both weights were tested on footage acquired by the two different camera types. A one-minute excerpt from both cameras covering the same area of the university mall was used for this purpose. Both excerpts were lowered to 1 frame per second at their original resolution, to allow for faster processing time and recording location datapoints at one snapshot per second. A total detection count was established by employing manual counting, combining all the detected pedestrians in each frame. For both excerpts, 360 counts were recorded which took 39-41 minutes to complete.

Afterwards, the YOLOv4 and self-trained detection weights were tested on both the excerpts under the Tensorflow Convolutional Neural Network (CNN) with the confidence score set to 50% and above. This filters out low probability detections. The YOLOv4 detection on the excerpt filmed by the DJI drone returned a total of 271 correct detections and 17 incorrect detections, while the self-trained weights returned 248 correct detections and 63 incorrect detections. On the footage excerpt from the Mi Dash Cam 2, the YOLOv4 returned 89 correct detections and 130 incorrect detections while the self-trained weight returned 249 correct detections and 404 incorrect detections. These detection processes took 40-57.5 seconds to complete (Table 1). Amongst the tested options, the YOLOv4 weight and the DJI drone combination performed the best with 75.3% of accuracy and fewer incorrect detections while YOLOv4 and the Mi Dash Cam 2 has the lowest detection accuracy. These differences revealed how the image resolution and colour balance affect the detection accuracy. The Mi Dash Cam 2 has a lower image resolution and quality compared to the DJI drone. Despite filming the same location, an area of plants was detected as 'potted plants' with YOLOv4 and the self-trained weights detected it as 'human' (Figure 4).

Table 1. Tracking accuracy data comparison

| **SETUP FOR DETECTION ACCURACY COMPARISON** | | | | |
|---|---|---|---|---|
| Detection Weights | YOLOv4 | Self- | YOLOv | Self- |
| Camera | DJI Phantom pro 4 | | Mi Dash Cam 2 | |
| Mean Average precision (mAP) (%) | 84 | 76 | 84 | 76 |
| Duration (min) | 1 | | 1 | |
| Frames per Second (fps) | 1 | 1 | 1 | 1 |
| Total Frame Counts | 60 | | 60 | |
| Resolution | 3840 x 2160 | | 2560 x 1600 | |
| **MANUAL PEOPLE COUNTING** | | | | |
| Total Detection Count | 360 | | 360 | |
| Processed Time (min) | 41 | | 39 | |
| **COMPUTER VISION OBJECT DETECTION (CVOD)** | | | | |
| Detection Class | 'Person' | 'Human' | 'Per- | 'Human' |
| Confidence Score Threshold (%) | 50 | 50 | 50 | 50 |
| Total Detection Count | 271 | 248 | 89 | 249 |
| Detection Accuracy (%) | 75.3 | 68.9 | 24.7 | 69.2 |
| Incorrect Detection | 17 | 63 | 130 | 404 |
| Processed Time (min) | 42.3 | 40 | 55.4 | 57.5 |

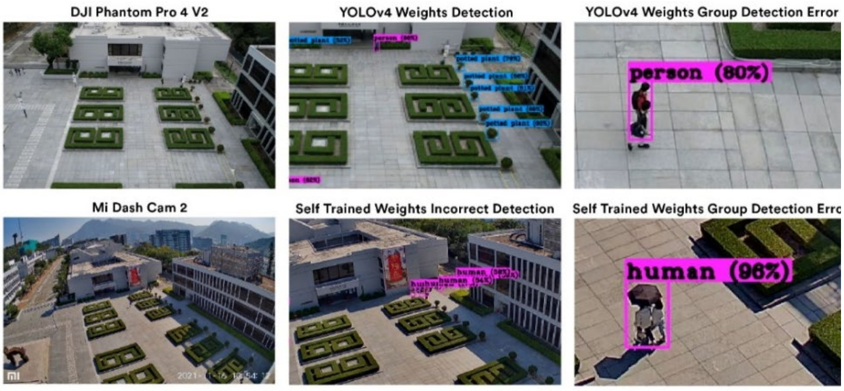
DJI Phantom Pro 4 V2
YOLOv4 Weights Detection
YOLOv4 Weights Group Detection Error
person (80%)
Mi Dash Cam 2
Self Trained Weights Incorrect Detection
Self Trained Weights Group Detection Errc
human (96%)

*Figure 4. Tracking bounding box and detection confidence value*

The low filming quality of the Mi Dash Cam 2 also directly impacted the performance of YOLOv4 as the COCO dataset that is utilized for the machine learning training process includes mostly close-up imageries of the human figure. However, the self-trained weights achieved a 69.2% detection accuracy on the Mi Dash Cam 2 footages, which could be further improved by enlarging the image data set that is being trained under the detection class 'human'. Both detection weights underperformed in the detection of people that walk in groups. Inclusion of extra detection classes would improve the quality of the detection data. At the current stage of development, the YOLOv4 and DJI drone footage combination was chosen for their accuracy.

## 2.5. DATA SPATIAL PROJECTION

The object tracking process returns a set of U,V coordinates per tracking ID that represents the pedestrian tracked in pixel space. The set of coordinates are specified as the centroid of the bounding box generated by YOLOv4. To turn the coordinates from pixel space to real world coordinates, the OpenCV function *getPerspectiveTransform* is used to perform the calculation of a matrix operation to project a set of points from one 2D plane to the map of the site generated in the photogrammetry process (Figure 5). The process requires a set of non-collinear coordinate points that represents the corners of a quadrilateral to be defined in a custom Python script. To obtain the coordinate points required for the transform process, four identical points between the video and the map were identified as pixel locations. The accuracy of these two sets of coordinate points is crucial to the perspective transform process. In selecting references, the points should be as far apart as possible.

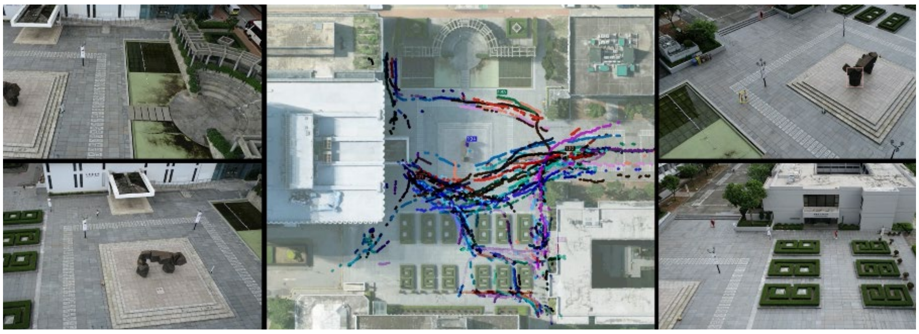

*Figure 5. Tracked trajectory of 1730-1738, 28/07/2021 (510 frames, 4575 detection points)*

The pixel-to-map perspective transformation performs only a linear translation of coordinates, it does not consider if a background pixel falls accurately on the 2D surface. Therefore, it is important to undistort the videos produced by the camcorder, which has 140 degrees field of view wide angle lens and produces a radial distortion that causes straight lines to appear curved. Using OpenCV's camera calibration function *cv.findChessboardCorners*, the image points of a checkerboard grid were located in ten reference photos taken with the camcorder. A distortion coefficient was then generated to undistort the whole data set. The result of the calculation outputs the tracker ID, timestamp, and map pixel coordinates data as a list in CSV format, which could then be used for analysis in the Rhino/Grasshopper environment.

## 2.6. CO-PRESENCE ANALYSIS

A subsequent analytical process was developed to interpret the pedestrian location data, evaluating “the in-between space that facilitates co-presence and regulates interpersonal relationships” (Madanipour, 2003, p. 206). In the analysis of our case study spaces, we evaluate various distances between people based on proxemic interactions theory (Hall, 1966), which describes how people "perceive, interpret, structure, and (often unconsciously) use the micro-space around them, and how this affects their interaction and communication with other nearby people" (Marquardt & Greenberg, 2015, p. 33). In our workflow, we employed a customised algorithm to analyse the closeness of individuals using the discrete proxemic zones defined by Hall: intimate (0 - 0.5 m), personal (0.5 – 1 m), social (1 – 4 m) and public (> 4 m) (Hall, 1966). The tool analyses the distance to all other people within the public space, and groups and counts people who are within the thresholds of social and personal space. Figure 6 illustrates this analysis, using a single time-frame snapshot observation, mapped on the 3D model of the site.

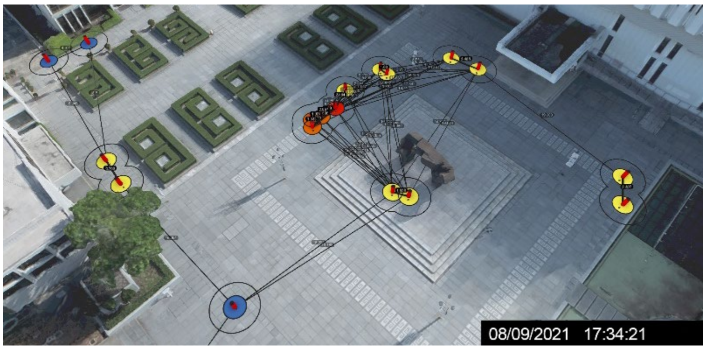

*Figure 6: Analysis of people locations and co-presence, based on a snapshot observation*

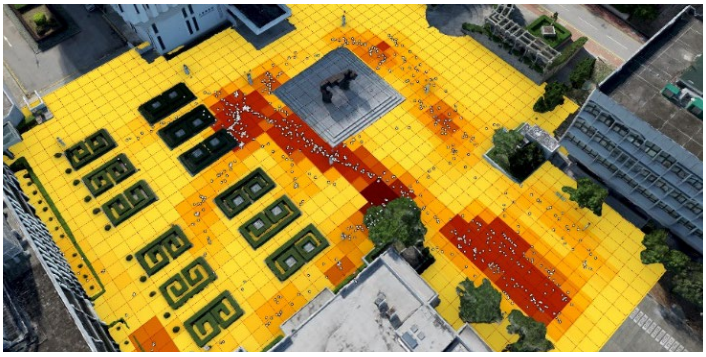

*Figure 7: Space-centric analysis of people locations and closeness, based on the combined pedestrian location data of 8.5 minutes of drone scanning*

In a final step of data translation, the human-centric analysis of people densities was translated into a space-centric analysis, defining the statistical occurrence of user presence and co-presence as a feature of the various locations within the case study space. The mapping process follows a basic logic of defining a spatial grid of cells, defined by their size; in this test, the grid spacing was set to two meters. The number of people location markers is counted within each cell, and the social proximity value is also recorded. For this analysis, multiple datasets relating to various time intervals can be combined, to produce insights into the general statistical patterns of space occupancy as they occur over longer periods of time. Figure 7 illustrates a data mapping of the entire time period of recording through the four static drone locations described in section 2.1 and Figure 2, compiled into one analytical visualisation projected onto the photogrammetry model.

## 3. Conclusions and Future Directions

This research has developed a methodology that employs UAV and building-mounted cameras, to construct a pedestrian Real-Time Locating System (RTLS), using Computer Vision Object Detection (CVOD) techniques to create digitalised pedestrian trajectory data. The methodology testing has demonstrated how multiple drone or building-mounted cameras can be used to capture different angles and segments of a public space, allowing for the method to be deployed in complex urban areas consisting of various spaces. The integration of multiple overlapping camera positions into a single analytical framework allows the methodology to be scalable across larger geographic areas, overcoming the limitations of traditional observational techniques for ethnographic study of space. Referring to concepts from humanistic geography and time geography, the methodology paves the way towards a system for automated detection of behavioural patterns related to social interaction and community forming, understanding specific behaviours such as avoiding, gathering, interacting, collaborating and dwelling in relation to specific environmental characteristics, public space morphology and configuration. Correlation analysis between behaviours and environmental attributes can produce insights in urban design principles that are conducive to socialising and placemaking, which would facilitate design guidelines for public spaces that support the social and mental health of individuals and communities.

As we will continue developing the methodology through case study applications, systematic analysis can reveal more detailed insights into which facilities are used more often, when and for how long, and how people move or interact around certain spaces. Public space design guidelines identified by urbanists such William Whyte and Jan Gehl will be able to be verified by data-driven and context-specific research, analysing interpersonal relationships, cultural norms and behaviour in relation to environment and context. The data-driven analysis of social dynamics can address important urban issues relating to safety and community when considering urban night-time settings, vulnerable people, and specific neighbourhood typologies such as old urban districts or public housing. It is our aim to continue developing research methodologies and digital visualisations of the social use of public spaces, to engage academics, policy makers and urban designers in conversations about improvements to existing, and the creation of new public spaces as part of the development of sustainable future cities.

## Acknowledgements

This project was supported by a Direct Grant for Research from the Social Science Panel, The Chinese University of Hong Kong, grant number 4052224.

## References

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning,* 215, 104217.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies,* 68, 285-299.

De Stefani, E., & Mondada, L. (2018). Encounters in public space: How acquainted versus unacquainted persons establish social and spatial arrangements. *Research on Language and Social Interaction,* 51(3), 248-270.

Friedmann, J. (2010). Place and place-making in cities: A global perspective. *Planning Theory & Practice,* 11(2), 149-165.

Gehl, J., & Svarre, B. (2013). *How to study public life.* Washington, DC: Island press.

Gou, Z., Xie, X., Lu, Y., & Khoshbakht, M. (2018). Quality of Life (QoL) Survey in Hong Kong: Understanding the Importance of Housing Environment and Needs of Residents from Different Housing Sectors. *International Journal of Environmental Research and Public Health*, 15(2), 219.

Hall, E.T., (1966). *The Hidden Dimension.* Doubleday, Garden City, NY.

Jacobs, J. (1961). *The Death and Life of Great American Cities.* New York: Vintage Books.

Low, S. (2016). *Spatializing Culture: The Ethnography of Space and Place*. London, England; New York, New York: Routledge.

Low, S., Simpson, T. and Scheld, S. (2019). *Toolkit for the Ethnographic Study of Space (TESS)*, Public Space Research Group, Center for Human Environments, The Graduate Center, City University of New York.

Madanipour, A. (2003). Social exclusion and space. In R. LeGates, and F. Stout, *The city reader* (Third ed., pp. 181-189). London: Routledge.

Marquardt, N., & Greenberg, S. (2015). Proxemic interactions: From theory to practice. *Synthesis Lectures on Human-Centered Informatics*, 8(1), 1-199.

Mehta, V. (2014). Evaluating public space. *Journal of Urban design*, 19(1), 53-88.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering,* 22(10), 1345-1359.

Schaumann, D., Kalay, Y.E., Hong, S. W., & Simeone, D. (2015). Simulating human behavior in not-yet built environments by means of event-based narratives. In *Proceedings of the symposium on simulation for architecture & urban design* (pp. 5-12).

Sidla, O., Lypetskyy, Y., Brandle, N., & Seer, S. (2006). Pedestrian Detection and Tracking for Counting Applications in Crowded Situations. *2006 IEEE International Conference on Video and Signal Based Surveillance*, 70.

Talen, E. (1999). Sense of community and neighbourhood form: An assessment of the social doctrine of new urbanism. *Urban studies,* 36(8), 1361-1379.

Tuan, Y. (1976). Humanistic Geography. *Annals of the Association of American Geographers,* 66(2), 266-276.

Westerink, J., Haase, D., Bauer, A., Ravetz, J., Jarrige, F., & Aalbers, C. B. E. M. (2013). Dealing with Sustainability Trade-Offs of the Compact City in Peri-Urban Planning Across European City Regions. *European Planning Studies*, 21(4), 473–497. https://doi.org/10.1080/09654313.2012.722927

Whyte, W. (1980). *The social life of small urban spaces.* Washington, D.C.: Conservation Foundation.

Williams, D. R. (2014). Making sense of 'place': Reflections on pluralism and positionality in place research. *Landscape and Urban Planning*, 131, 74-82.

Zakariya, K., Harun, N. Z., & Mansor, M. (2014). Spatial characteristics of urban square and sociability: A review of the City Square, Melbourne. *Procedia-Social and Behavioral Sciences*, 153, 678-688.