

Classifying Human Activities in Urban Spaces with a Multimodal AI: Towards a Massive Assessment of Urban Affordances

Ivan Blečić[1(✉)], Alessandro Floris[1], Giulia Giliberto[1], and Giuseppe A. Trunfio[2]

[1] Department of Civil and Environmental Engineering and Architecture, University of Cagliari, Cagliari, Italy
{ivanblecic,a.floris62,giulia.giliberto}@unica.it
[2] Department of Biomedical Sciences, University of Sassari, Sassari, Italy
trunfio@uniss.it

Abstract. We present a tool that leverages a Multimodal Large Language Model (MLLM) for the automatic classification of human activities from images of urban scenes. Starting from an image of spaces populated with people, the tool is capable to classify them according to five features: age group, sex, bodily posture, activity level and social configuration. The tool implements a sequential pipeline consisting of Faster R-CNN for person detection, followed by postprocessing and two consecutive applications of GPT-4o models for refined image description and information extraction.

In the paper we also present an experimental test used to preliminary validation of the tool, comparing the ground truth on 24 images of urban scenes with the estimates provided by the tool, yielding a good degree of alignment.

The tool is part of the wider research programme of massive assessment of urban affordances, within the framework of the capability approach.

Keywords: urban affordances · capability approach · urban space analysis · Multimodal Large Language Model · urban evaluation modelling

1 Introduction

Public spaces play an important role in shaping human experiences, social interactions, and people's capabilities [1]. Understanding how people use and interact with urban spaces is fundamental for designing inclusive and enabling environments [2–4].

Since the 1960s, pioneers such as Jane Jacobs [5], William H. Whyte [6], and Jan Gehl [7, 8] highlighted the importance of creating urban spaces adopting a human-centred perspective, rather than overly relying on formal "aesthetic" forms or technical solutions for the built environment.

Building on these foundational ideas, there has been a growing demand for methods identifying relationship between people and their urban environments, aiming to better understand how urban spaces are used in everyday life, and to provide insights to inform policies and guide the physical transformation of cities.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15890, pp. 341–357, 2026.
https://doi.org/10.1007/978-3-031-97606-3_23

Empirical methods within this domain often employ direct observation as a tool for surveying, mapping, and categorising how urban spaces are used by people [9].

Traditionally, such observational studies on the use of public spaces have been conducted through direct human observation. While this approach offers valuable qualitative insights, it is inherently time-consuming, difficult to scale up, and possibly subject to observer bias. In response to such challenges, recent advancements in machine learning and artificial intelligence provide new opportunities for automating and enhancing behavioural analysis in urban contexts [10]. Indeed, the past decade has seen a growing adoption of computer vision and machine learning techniques for automating behavioural analysis in urban contexts.

In this paper we present one such tool, leveraging a Multimodal Large Language Model (MLLM), for the automatic classification of human activities from images of scenes of urban spaces. Specifically, starting from an image of urban space populated with people, the tool is capable to classify them according to five features: age group, sex, bodily posture, activity level and social configuration.

As we will more extensively present in the Background section, besides automatising and hence allowing for massive assessment of activities in urban spaces, the tool is instrumental for the development of a wider research programme of the systematic assessment of urban capabilities [11–14], within which a relevant role is played by the physical affordances [15] of urban space, conceptualised as a subset of urban opportunities which are directly perceivable and actionable in the urban physical environment.

The paper is organised as follows: the following Sect. 2 presents some relevant background both on the theoretical and methodological framework, as well as on the techniques and technologies employed for the development on the tool; Sect. 3 provides more details on the observation protocol and technical details on the functioning of the tool; in Sect. 4 we present the experimental test and discuss its results; finally, Sect. 5 offers some concluding remarks and perspectives for future development.

The present study was carried out within the TERSICORE (Technologies for Evaluation and Research for Sustainability Innovation: a Community-Oriented Regenerative Evolutions) research project, the aim of which is to develop a set of protocols for the assessment of the psycho-social dimension of urban space in order to support projects of urban regeneration and to evaluate their impacts.

2 Background

This background section is subdivided into two subsections. The first is dedicated to theoretical and methodological framework liveable which to situate the tool for automatic assessment of human activities in urban spaces. The second subsection presents background on techniques and technologies that were employed for the development of the tool.

2.1 Urban Capabilities, Opportunities and Affordances

In cities people engage in activities, they perform different behaviours in different urban spaces. These activities are characterised by the interaction of people with objects, other

people, and the environment, with the intention of achieving goals and the attainment of certain functionings.

The term "functioning" is here used specifically in the context of the capability approach [16, 17], in which a human functioning is attained among those within individual's set of capabilities.

A capability may be defined as a person's substantive freedom to choose among various options of doing and being that they have reason to value. A capability presupposes two constitutive elements: (1) individual ability, understood as the internal capacity—possessed though not necessarily exercised—to act or to be in particular ways; and (2) opportunity, referring to the external conditions that render the exercise of such capacity feasible. A person may be said to possess a given capability only when both components—internal and external, ability and opportunity—are concurrently present. The physical configuration of urban space—the material infrastructure of the city—affects capabilities primarily through its impact on the opportunity component.

The opportunities relevant to attain urban capabilities may in principle be a broad set of circumstances, including institutional, economic, social, and spatial conditions. Within these, the *affordances* in Gibson's sense [15] may be conceptualised as a subset of opportunities, in particular those that are directly perceivable and actionable in the urban physical environment.

Hence, the functionings attained within their capability set comprise activities people perform within urban space, so these activities are the "mediating factor" between the people and the physical urban environment. Concrete actions are goal-oriented, driven, and adjusted to the circumstances. Even while a goal and a reason might not change, the approach used to get there might, depending on the situation and on the environment [18]. And this possibility of action hinges critically on the available affordances of urban space.

While the objective of a massive assessment of affordances of urban spaces provides the general drive to our research programme, the assessment tool presented here is just a stepping stone to that more general aim. Instead of actually assessing affordances of physical spaces, it performs a more limited task of assessing people's activities, which are directly observable. In other words, it observes actual spatial functionings taking place in urban spaces.

To this end, behavioural mapping represents a methodological approach based on structured observation, combining various detection, mapping, and quantification techniques of activities performed at different times by people occupying a defined space. This method enables the identification of patterns in space utilisation, social interactions, and activity distributions [19].

From a human-centred perspective, the approach helps explore both the quantity and quality of the various activities performed by people who move, stay, play, and spend their time in different ways. The data collection process through observation, analysis, and visualisation can highlight the presence of different social groups in the study area and their behaviour and thus of the affordances relating to that public space. The information gathered about practices in a particular urban context potentially describes the general framework in terms of socialisation, time budgets, and physical activity. The results, therefore, provide an understanding of user dynamics in specific places [20].

So, the objective of urban analysis is twofold: it is both an observational and a documental process, and it should also lead to positive change in the built environment [21]. The outcomes can serve as inputs for formulating urban regeneration programs and for evaluating the effectiveness of transformative interventions.

2.2 AI-Based Tools: Techniques and Technologies

In recent times, conventional observational methods such as direct observation and enumeration are being complemented by technological advancements including GPS, GIS, skin sensors, remote sensing, and closed-circuit television (CCTV), amongst others. The rapid advancements in computer processing speed and power have rendered it feasible to integrate diverse sensing and observation modalities in ways that were hitherto inconceivable.

Many recent studies have focused on person detection, employing *object detection* models to localise individuals in images or video streams. For example, in [22], the authors used an advanced variant of the YOLO (You Only Look Once) series of real-time object detectors to identify individuals in a public square and classify them according to demographic and behavioural attributes, achieving approximately 83% accuracy across categories such as walking, sitting, and exercising. A complementary approach is presented in [23], where the authors employed YOLOv4 [24] for person detection and CNN-based classifiers (*ResNet-18* [25] and *EfficientNet* [26]) for gender classification, trained on a custom-annotated dataset. Rather than directly recognising physical activities or postures, their method inferred usage patterns in an urban park based on the spatial distribution, gender, and temporal presence of detected individuals. This enabled the identification of behavioural trends—such as which demographics utilised specific areas of the park—without the need for explicit action or posture recognition.

Another line of research gaining traction in public space analysis involves *pose estimation*, which enables the automatic classification of bodily postures (e.g., upright, seated, recumbent) by extracting and analysing skeletal keypoints. However, deploying such techniques in uncontrolled outdoor environments remains challenging due to factors such as occlusion, variable lighting, and background complexity. As highlighted by Cormier et al. in [27], the lack of large-scale datasets tailored for public surveillance settings has hindered the development of generalisable pose estimation models, and there is a continued need for lightweight systems capable of operating in real-time with limited computational resources.

A key innovation in the field has come with the emergence of Multimodal Large Language Models (MLLMs)—extensions of Large Language Models (LLMs), which are deep neural networks trained on vast corpora of text to generate and understand human language. While LLMs, operate solely on textual input, MLLMs are designed to process and reason over multiple modalities, particularly combining visual and textual inputs, allowing them to interpret images and extract structured semantic information in a zero-shot fashion, that is without requiring task-specific training. This enables scene understanding, extracting semantic and structured information from visual content without the need for task-specific training. An early application in the context of urban analysis is found in [28], who used a MLLM to assess walkability from Google Street View imagery, demonstrating the ability of these models in multimodal reasoning. Also,

[29] explored the use of GPT-4 Vision in the automated analysis of urban morphology and public space usage using both Google Street View and CCTV imagery. Their approach consisted of prompting the model to generate structured JSON outputs describing elements such as the presence and activities of people, built environment quality, and public safety indicators. To address inconsistencies and hallucinations—such as incorrect detection of "danger" from suggestive elements like traffic cones, or fluctuating labelling of features like "mixed-use areas"—they submitted each image five times and aggregated the outputs to improve reliability. Despite these efforts, the authors emphasise that the system's outputs should be treated as "suspected" rather than definitive detections, requiring human verification and not suitable for critical decision-making.

These limitations point to a key challenge in applying general-purpose MLLMs directly for fine-grained urban behavioural analysis: while they are flexible and powerful, their outputs can be inconsistent, ambiguous, or overly sensitive to context. Our approach follows this recent line of research leveraging MLLMs and LLMs for urban scene understanding, explicitly addressing these issues by introducing a structured pipeline designed to enhance robustness and interpretability. Specifically, it extracts consistent and reliable information on individuals' bodily posture, activity level, social configuration, age range, and sex from images captured in public spaces.

After extensive experimentation with several multimodal strategies—including LLaVA [30], Gemini [31], and GPT-4o [32]—we identified the most effective approach as a sequential pipeline consisting of Faster R-CNN [33] for person detection, followed by postprocessing and two consecutive applications of GPT-4o models for refined image description and information extraction. The details of our approach are presented below in Sect. 3.1.

3 Observation Protocol

The presented AI tool is based on and aims at reproducing an observation protocol designed to be performed by human auditors/observers. The protocol, more extensively presented in [34], has the purpose to gather information about what people do in public spaces, and has been tested in real-world contexts.

The observation protocol was designed to control for different variables, such as time of day, day or week, season, location and so on. A controlled observation follows a clear set of steps. These steps include what, where, when, who, how and how long to observe.

In relation to the object of study, i.e. people's actions in open spaces, the observation is quantitative as it generates numerical outputs of frequency counts.

The research method is founded upon a series of systematic questions, enabling the articulation of diverse activities undertaken by various groups of people across multiple levels. Specifically, context-specific questions are derived from the integration of five fundamental questions [19]:

- *How many people?* – The first question concerns the number of people in relation to the space, i.e. how many people are present in the observed scene.
- *Who?* – The second question aims to categorise the actors in the space, dividing them into different groups according to two variables:

  - *age*, classified on a generational basis (child, adult, senior);
  - *gender* (male, female).

– *Where do people carry out their activities?* – The third question focuses on the activities people perform in space, with a particular emphasis on the mapping of locations that offer opportunities for action. The data collected enables the description of the positioning of individuals according to the activities they perform and the elements present.
– *What activities take place?* – The fourth question explores the nature of the activity, with respect to a specific taxonomy of activities, subdivided in three features:

  - *body posture*, describing physical position or bodily configuration of the person, classified into "upright" (standing, leaning, squatting), "seated" (sitting on a bench, ground, ledge, stairs, wall, etc.), and "recumbent" (lying down, lounging);
  - *physical activity*, movement, or exertion; the classes are "sedentary" (stationary, minimal movement while standing/sitting/lying, for example resting, reading, eating, playing cards or a board game), "moderate" (walking, slow ambulatory movement), and "vigorous" (running, jumping, playing, exercising);
  - *social configuration*, indicating the social context in which the individual is situated; the classes are "solitary" (alone, no direct social interaction, solitary presence), "dyadic" (in a pair, interacting with one other person), and "group" (in a group of three or more, engaged socially).

– *How long does the activity last?* – The fifth question considers the temporal component associated with activities performed in public space. Specifically, it assesses the duration of the activities. Duration can provide insights into the quality of the spaces.

3.1 The Tool: A Multimodal AI Pipeline for Automatic Observation of Activities in Urban Spaces

The purpose on the tool presented here is to implement an AI pipeline for automatic observation of activities in urban spaces, based on the observation protocol briefly presented in the previous subsection. The tool accepts images of urban scenes populated with people, and automatically classifies people in the scene according to five features: age group (child, adult, senior), sex (female/male), bodily posture (upright, seated, recumbent), activity level (sedentary, moderate, vigorous) and social configuration (solitary, dyadic, group).

After extensive experimentation with several multimodal strategies—including LLaVA [30], Gemini [31], and GPT-4o [32]—we identified the most effective approach as a sequential pipeline consisting of Faster R-CNN [33] for person detection, followed by some postprocessing detailed below and two consecutive applications of GPT-4o models for refined image description and information extraction.

A schematic overview of this workflow is shown in Fig. 2, which visually represents each step from person detection to the final structured output.

The pipeline initiates with the detection of individuals within the image, for which we used Faster R-CNN, from Detectron2. Faster R-CNN is an advanced object detection model utilising a Region Proposal Network (RPN) to identify potential bounding boxes that might contain objects. The adopted version of the model employs a ResNeXt-101 backbone architecture, known for its superior performance in extracting rich visual features, and Feature Pyramid Networks (FPN), which enhance detection accuracy across multiple scales by leveraging features at various resolution levels. The model is trained for extended epochs (3×), improving its robustness and accuracy in detecting complex objects, especially humans, across diverse environments.

Detectron2 is a state-of-the-art, open-source platform developed by Facebook AI Research (FAIR) for building, training, and deploying object detection, segmentation, and other visual perception models. It provides modular, highly customisable components and pre-trained models that facilitate rapid experimentation and deployment of advanced computer vision applications. Note that, in views of the subsequent steps, identified bounding boxes undergo padding to ensure complete encapsulation of the detected individuals. In the present application, padding values are calculated proportionally, extending each bounding box by 40% horizontally and 20% vertically, thus providing sufficient contextual visual information for following analysis.

As shown in Fig. 1, to reduce the complexity of the visual input passed later to the language model, bounding boxes intersecting or closely positioned are clustered into singular bounding boxes.

This clustering employs an iterative approach, merging overlapping bounding boxes until no further merges are possible, significantly enhancing the coherence and clarity of subsequent image cropping and analysis. Importantly, this step does not necessarily isolate single individuals or single groups of interacting people but rather aims to ensure that cropped regions passed to the language model do not contain an excessive number of distinct people. In fact, overcrowded crops tend to reduce the precision of the multimodal language model's descriptive capacity. Clustering therefore plays a crucial role in optimising the quality of the generated image descriptions, preserving a balance between sufficient context and visual clarity.

Subsequently, the clustered bounding boxes are used to crop sub-images from the original frame. Each cropped image is then independently analysed by a MLLM, which was GPT-4o in the present application, to produce detailed descriptions, capturing essential visual attributes and contextual cues. GPT-4o's proved quite robust in vision understanding and could ensure precise interpretations of visual content, addressing even partially visible individuals.

Generated textual descriptions from the MLLM are further processed through a secondary LLM pass to extract structured information systematically. For this step also we used GPT-4o after some experimentation with alternative models. This step assigns definitive values for bodily posture (upright, seated, recumbent), activity level (sedentary, moderate, vigorous), social configuration (solitary, dyadic, group), age range (child, adult, senior), and sex (male, female) for each individual, even under conditions of visual uncertainty.

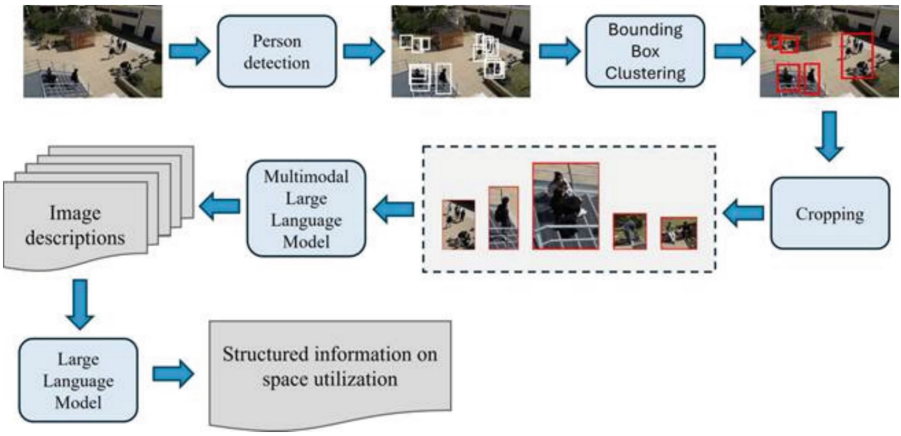


Fig. 1. Multimodal AI pipeline for structured observation of public spaces. Individuals are detected with Faster R-CNN and grouped by bounding box overlap. Cropped regions are expanded and analysed by GPT-4o, which generates descriptions later parsed to extract posture, activity level, social configuration, age group, and gender.

Extracted data are compiled into structured JSON files, facilitating downstream statistical analysis and comparative evaluations. These files comprehensively document each individual's characteristics and the total count of visible persons per analysed image.

The approach described above effectively mitigates several issues reported in [29], including the inconsistency of outputs and hallucination of scene elements. In contrast to their single-step prompting method, our two-stage pipeline decouples image interpretation from structured data extraction. This reduces ambiguity, ensures more accurate attribute labelling, and facilitates validation against observational data. Furthermore, the use of bounding box clustering and context padding contributes to increased robustness when analysing crowded scenes or partially occluded individuals.

The following sections present a comparative evaluation of this pipeline against manually annotated data, to assess its accuracy and reliability in real-world scenes.

4 Experimental Test

To test the tool, we have collected images from 24 different scenes across 5 open urban public spaces in the city of Cagliari (Italy). The images were captured under favourable weather conditions and during daylight hours over the course of three days in March 2025.

Each image was classified by a human auditor according to five features:

- age group ("child", "adult", "senior");
- sex ("female" or "male");
- body posture ("upright or "seated"); the "recumbent" category from the original protocol was not present in the sample;

- activity level ("sedentary" or "moving"); "moving" category comprised of only "moderate" activity levels, since "vigorous" level was not present in the sample;
- social configuration ("solitary" or "group"); the "group" category aggregates "dyadic" and "group" categories from the original protocol.

Figure 2 presents an example of the procedure applied for the manual evaluation of a sample images. The images were then submitted to the classification algorithm, providing the estimate classifications for the same four features.

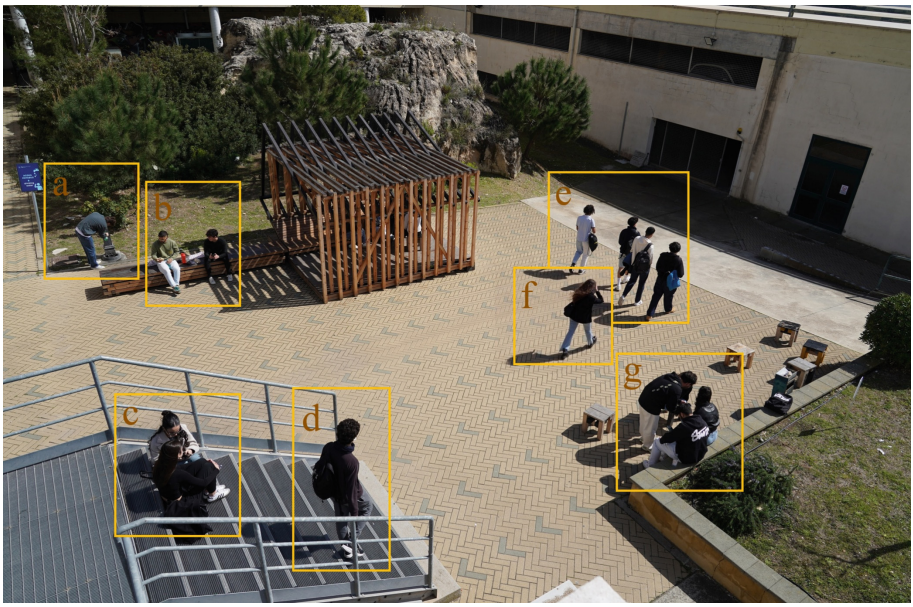
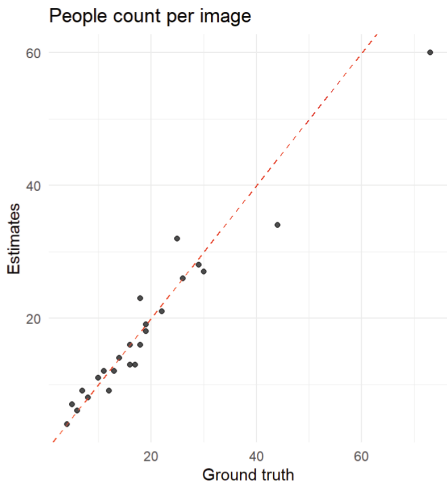


Fig. 2. Example of manual evaluation of the captured scenes. The image presents seven social configurations (a–g), within which individual components are observed. *Configuration a* (solitary): 1 male, upright, sedentary; *configuration b* (group): 2 males, seated, sedentary; *configuration c* (group): 2 females, seated, sedentary; *configuration d* (solitary): 1 male, upright, moving; *configuration e* (group): 4 males, upright, moving; *configuration f* (solitary): 1 female, upright, moving; *configuration g* (group): [1 male, seated, sedentary], [1 male, upright, sedentary], [1 female, seated, sedentary].

The overall count of people in the two samples, ground truth and estimates, together with essential descriptive statistics, are summarised in Fig. 3.



Statistics	Ground truth	Estimates
Total count of people	462	438
Min	4	4
Median	16.50	15.00
Mean	19.25	18.25
Max	73	60

Fig. 3. People count: scatterplot (left); descriptive statistics (right)

The figures in the following pages – from Figs. 4, 5, 6 and 7 – show scatterplots and Bland-Altman plots, reporting the counts per each feature's category with respect to the two samples (ground truth and estimates).

To assess the alignment and distributional similarity between the ground truth and the estimates, several per-category error metrics and global divergence measures are calculated on relative data (i.e. percentages), comparing estimated and ground-truth category distributions.

MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) quantify the average absolute and squared deviations, respectively, between estimated and true proportions for each category. Bias indicates the mean signed error, capturing over- or under-estimation.

As global metrics, Chi-square distance, Kullback–Leibler (KL) divergence and Jensen–Shannon (JS) divergence are calculated for each of the four features. Chi-square distance measures the average relative squared difference between estimated and observed proportions, normalised by the observed values. Kullback–Leibler (KL) divergence quantifies the information loss when using the estimated distribution to approximate the ground truth. Jensen–Shannon (JS) divergence is a symmetric and smoothed version of KL, bounded between 0 and 1.

For the three global metrics, we estimated the confidence intervals using a non-parametric bootstrap resampling procedure. Specifically, we performed 1,000 bootstrap iterations, in each of which the set of samples (i.e., per-image predictions) was resampled with replacement, and the target metrics (i.e., Chi-square distance, KL divergence, Jensen-Shannon divergence) were recomputed. The 95% confidence interval was then computed as the 2.5th and 97.5th percentiles of the resulting bootstrap distribution.

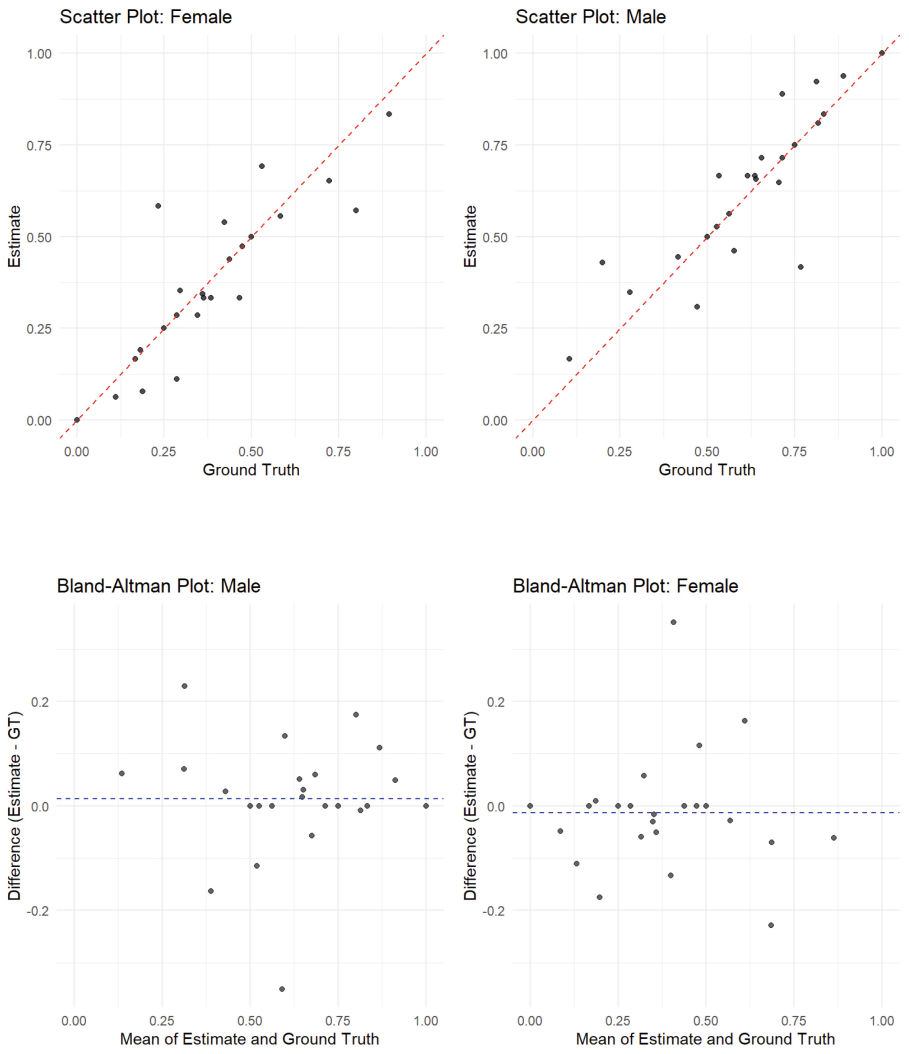


Fig. 4. Feature: Sex (female or male); values expressed in percentages (%).

All the metrics and confidence intervals are reported in Table 1. In the table, we report values just for one category within each feature, given that all the features hold only binary categories (e.g. female/male, upright/seated, etc.), so the values reported for MAE and RMSE are identical for both categories of each feature, while the Bias is sign-inverted (e.g. for the "Sex" feature, the estimate bias for "female" −.0132, hence the estimate bias for "male" is .0132).

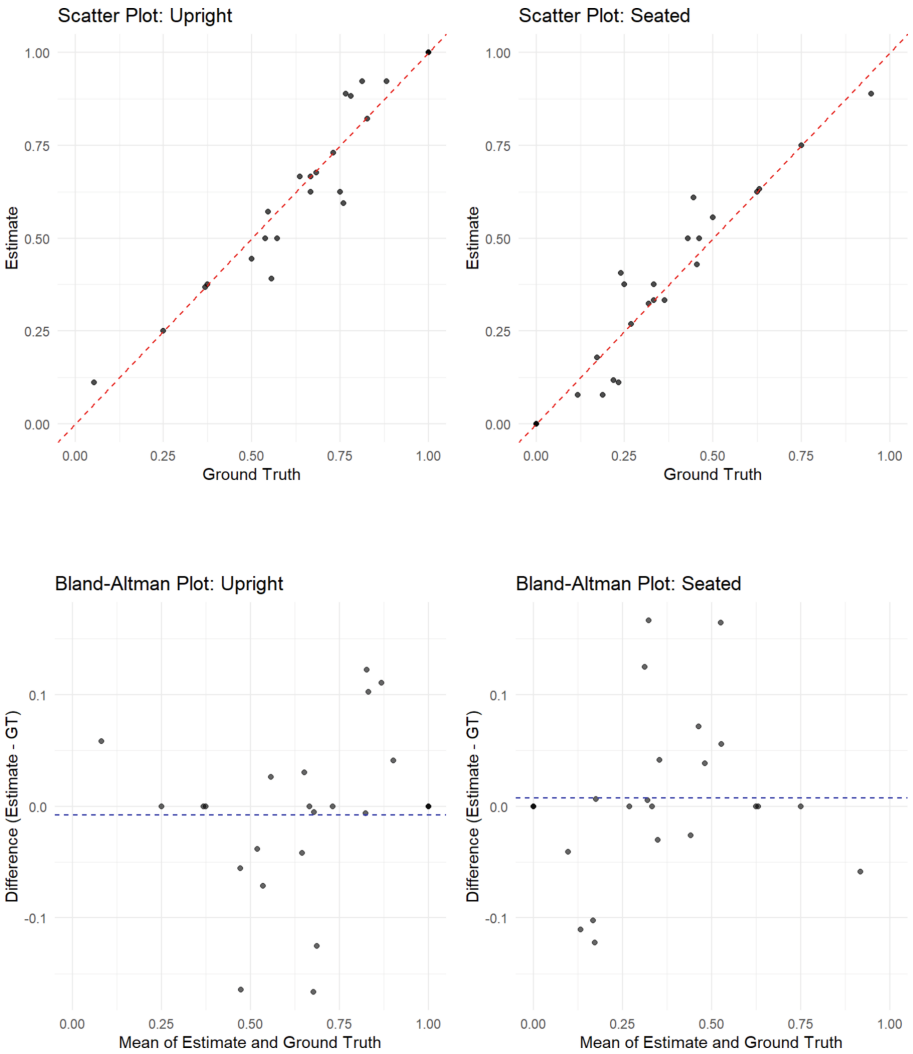


Fig. 5. Feature: Body posture (upright or seated); values expressed in percentages (%).

4.1 Discussion

The per-category error metrics (MAE, RMSE, Bias), as well as the global divergence metrics reveal satisfactory, although not uniform degrees of alignment between estimated and ground-truth distributions. The lowest divergence values were observed for Sex (Chi-Square = .07, KL = .03, JS = .01) and Body Posture (Chi-Square = .03, KL = .02, JS = .00), indicating high consistency between algorithmic estimates and actual category proportions for these features.

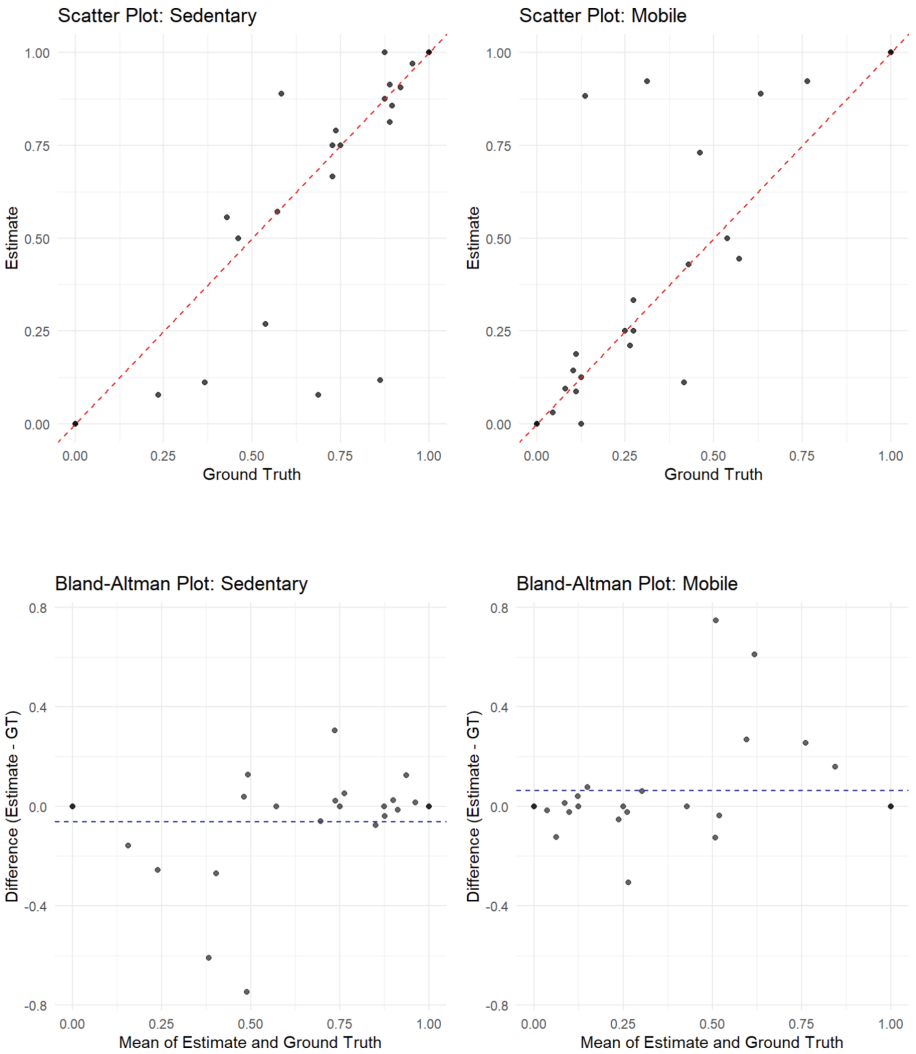


Fig. 6. Feature: Activity level (sedentary or mobile); values expressed in percentages (%).

In contrast, Activity Level exhibited higher divergence across all three measures (Chi-Square = .33, KL = .20, JS = .03), suggesting that the classification algorithm performed somewhat worse with estimating the distribution of sedentary vs. mobile individuals. Social Configuration showed somewhat intermediate divergence (Chi-Square = .10, KL = .19, JS = .02), indicating some discrepancies, though less pronounced than for Activity Level.

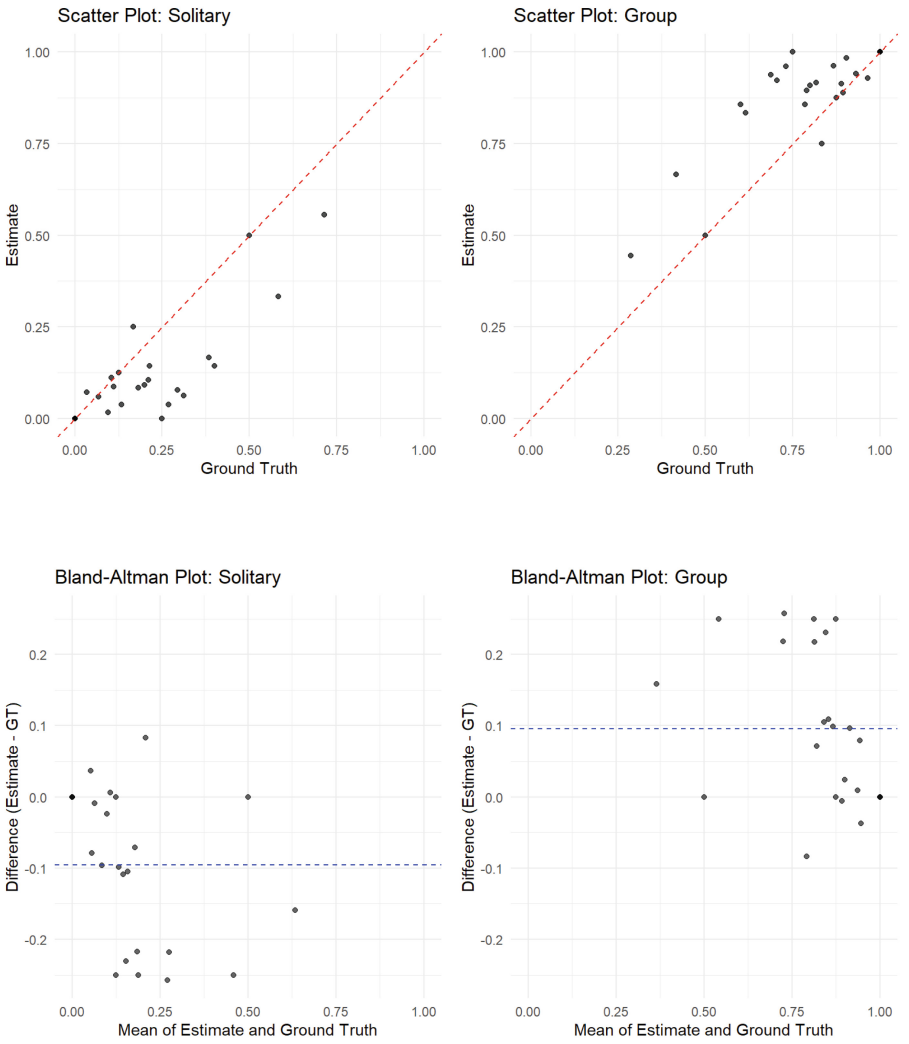


Fig. 7. Feature: Social configuration (solitary or group); values expressed in percentages (%).

Table 1. Per-category error metrics and global divergence measures comparing ground-truth and estimated category distributions. MAE: Mean Absolute Error; RMSE: Root Mean Squared Error; CSq: Chi-square distance; KL: Kullback–Leibler divergence; JS: Jensen–Shannon divergence; CSq | KL | JS CI: respective 95% confidence intervals.

Feature	MAE	RMSE	Bias	CSq	CSq CI	KL	KL CI	JS	JS CI
Sex (female/male)	.0711	.1114	− .0132	.07	.02-.14	.03	.01-.06	.01	.00-.01
Body posture (upright/seated)	.0485	− .0725	− .0076	.03	.01-.05	.02	.00-.03	.00	.00-.01
Activity level (sedentary/mobile)	.1226	.2268	− .0633	.33	.04-.84	.20	.05-.38	.03	.01-.06
Social configuration (solo/group)	.1063	.1428	− .0958	.10	.06-.15	.19	.05-.45	.02	.01-.03

5 Conclusions

The results presented suggest that the algorithm performs very well in estimating distributions for sex and posture, while further refinement may be needed for activity level and social configuration, although also in those cases the performance of the classification algorithm is fairly acceptable.

The relative greater difficulty to estimate the activity level (sedentary vs. mobile) is likely inherent to the nature of the input data, i.e. still images of urban scenes. In these, the mobile or sedentary status of people can often be intrinsically ambiguous, add hence hard to interpret or altogether impossible to discern. Should a child goalkeeper playing football but standing still in the image be classified as sedentary or as performing a vigorous physical activity? Likely, to improve the classification performance here it would be necessary to submit a video clip, or a sequence of frames shot in close time succession to better assess the dynamics of the activity.

Both these paths lie before us, rich with potential for further development of the tool.

Acknowledgments. This study was funded by:

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Qi, J., Mazumdar, S., Vasconcelos, A.C.: Understanding the relationship between urban public space and social cohesion: a systematic review. Int. J. Com. WB. **7**, 155–212 (2024). https://doi.org/10.1007/s42413-024-00204-5
2. Mitchell Tada, G.: Socially just public spaces are crucial to flourishing societies (2023)
3. Low, S.: Why Public Space Matters. Oxford University Press, Oxford (2022)
4. Tigran, H., Littke, H., Elahe, K.: Urban form and human behavior in context of livable cities and their public realms. SJPBS **3**(4), 325–339 (2020). https://doi.org/10.32474/SJPBS.2020.03.000167
5. Jacobs, J.: Death and Life of Great American Cities. Random House, New York (1961)

6. Whyte, W.H.: The social life of small urban spaces. In: Project for Public Space, New York (1980)
7. Gehl, J.: Life Between Buildings: Using Public Space. Island Press, Washington, DC (1971)
8. Gehl, J.: Cities for People. Island Press, Washington, DC (2010)
9. Szczepańska, A., Pietrzyk, K.: An evaluation of public spaces with the use of direct and remote methods. Land **9**, 419 (2020). https://doi.org/10.3390/land9110419
10. Weng, Q., et al.: How will ai transform urban observing, sensing, imaging, and mapping? NPJ Urban Sustain. **4**, 1–9 (2024). https://doi.org/10.1038/s42949-024-00188-3
11. Blečić, I., Cecchini, A., Talu, V.: The capability approach in urban quality of life and urban policies: towards a conceptual framework. In: Serreli, S. (ed.) City Project and Public Space. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-6037-0_17
12. Blečić, I., Cecchini, A., Congiu, T., Fancello, G., Talu, V., Trunfio, G.A.: Operationalising urban capabilities: the case of walkability. In: Proceedings of the 2015 HDCA Conference-Capabilities on the Move: Mobility and Aspirations, Washington, September 10, p. 2015 (2015)
13. Blečić, I., Cecchini, A., Talu, V.: Capability approach and urban planning. Fertile urban capabilities and quality of urban life of the most disadvantaged inhabitants. Archivio di Studi Urbani e Regionali **48**, 34–52 (2018). https://doi.org/10.3280/ASUR2018-122003
14. Blečić, I., Cecchini, A., Congiu, T., Fancello, G., Talu, V., Trunfio, G.A.: Capability-wise walkability evaluation as an indicator of urban peripherality. Environ. Plann. B Urban Anal. City Sci. **48**, 895–911 (2021). https://doi.org/10.1177/2399808320908294
15. Gibson, J.: The Ecological Approach to Visual Perception. Psychology Press, New York (1979)
16. Sen, A.: Development as Freedom. Oxford University Press, Oxford (1999)
17. Sen, A.: The Idea of Justice. Harvard University Press, Harvard, MA (2009)
18. Bærentsen, K.B., Trettvik, J.: An activity theory approach to affordance. In: Nordichi 2002. Proceedings of the Second Nordic Conference on Human-Computer Interaction, pp. 51–60 (2002)
19. Gehl, J., Svarre, B.: How to Study Public Life. Island Press, Washington, DC (2013)
20. Byrne, J.A.: Observation for data collection in urban studies and urban analysis. In: Baum, S. (ed.) Methods in Urban Analysis, pp. 127–149. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1677-8_8
21. Ye, Y., Li, D., Liu, X.: How block density and typology affect urban vitality: an exploratory analysis in Shenzhen. China. Urban Geogr. **39**, 1–22 (2017). https://doi.org/10.1080/02723638.2017.1381536
22. Wang, L., He, W.: Analysis of community outdoor public spaces based on computer vision behavior detection algorithm. Appl. Sci. **13**, 10922 (2023). https://doi.org/10.3390/app131910922
23. Gravitz-Sela, S., Levy, A., Zehavi, S., Bryt, O., Shach-Pinsly, D., Plaut, P.: Analyzing usage patterns from video data through deep learning: the case of an urban park. Comput. Environ. Urban Syst. **117** (2025). https://doi.org/10.1016/j.compenvurbsys.2024.102229
24. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
26. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2020)
27. Cormier, M., Clepe, A., Specker, A., Beyerer, J.: Where are we with human pose estimation in real-world surveillance? In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). pp. 591–601. Waikoloa, HI, USA (2022)

28. Blečić, I., Saiu, V., A. Trunfio, G.: Enhancing urban walkability assessment with multimodal large language models. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D.,C. Rocha, A.M.A., Faginas Lago, M.N. (eds.) Computational Science and Its Applications – ICCSA 2024 Workshops, pp. 394–411. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-65282-0_26
29. Komar, T., James, P.: LLM-vision in enhancing the understanding of public spaces. Abstr. ICA. **8**, 1–4 (2024). https://doi.org/10.5194/ica-abs-8-15-2024
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
31. Gemini Team Google: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2024)
32. OpenAI: GPT-4°. https://platform.openai.com/docs/models/gpt-4o
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
34. Manunza, A., et al.: Build it and they will stay: assessing the social impact of self-build practices in urban regeneration. Urban Sci. **9**(2) (2025). https://doi.org/10.3390/urbansci9020030