Towards Data-Informed Interventions: Opportunities and Challenges of Street-level Multimodal Sensing

Joao Rulff[1], **Giancarlo Pereira**[1], **Marcos Lage**[2],
Maryam Hosseini[3], **Claudio Silva**[1]

[1] New York University, [2] Universidade Federal Fluminense, [3] University of California, Berkeley

Figure 1. This figure highlights situations where pedestrians with limited mobility levels cross streets (A, C, D). These examples were identified using open-vocabulary detection models. (B) shows how street-level videos can contain valuable information regarding traffic lights.

Abstract

Over the past decades, improvements in data collection hardware coupled with novel artificial intelligence algorithms have made it possible for researchers to understand urban environments at an unprecedented scale. From local interactions between actors to city-wide infrastructural problems, this new data-driven approach enables a more informed and trustworthy decision-making process aiming at transforming cities into safer and more equitable places for living. This new moment unfolded new opportunities to understand various phenomena that directly impact how accessible cities are to heterogeneous populations. Specifically, sensing localized physical interactions among actors under different scenarios can drive substantial interventions in urban environments to make them safer for all. In this manuscript, we list opportunities and associated challenges to leverage street-level multimodal sensing data to empower domain experts in making more informed decisions and, ultimately, supporting a data-informed policy-making framework. The challenges presented here can motivate research in different areas, such as computer vision and human-computer interaction, to support cities in growing more sustainably.

1. Introduction

In the complex system of urban public spaces, diverse actors, i.e., pedestrians, cyclists, and drivers, interact in dynamic ways. Understanding these interactions has long been a goal of various fields ranging from transportation to social sciences. Computer vision methods opened new horizons for fine-level understanding of these dynamics by allowing scalable analysis of multimedia datasets. However, accessibility-related challenges have often been overlooked. Assistive devices such as wheelchairs, walkers, and canes, which are crucial to many urban dwellers, have not been well represented in standard benchmarks like semantic segmentation or object detection. This oversight leaves a significant gap in how urban environments are understood and designed for inclusivity. Recent developments, however, offer promising new avenues to address these gaps. Open-vocabulary object detection (OVD) models are emerging as

a powerful tool to enhance inclusivity in computer vision tasks. Unlike traditional models limited by predefined categories, OVDs have the potential to detect and describe a far broader range of objects, including assistive devices. As represented in Figure 1, these models can quickly identify a set of cases where low-mobility pedestrians are interacting with urban intersections.

In tandem, high-performance models for action recognition and pose estimation now allow researchers to track the behaviors and interactions of pedestrians with remarkable detail. These models, capable of identifying actions like walking, sitting, and running, offer critical insights into how pedestrians navigate the urban landscape, especially in scenarios where they interact with other actors, such as cyclists or vehicles. Multimodal data, including machine listening models, further enrich our understanding of urban dynamics by providing complementary information such as sound cues. Acoustic localization algorithms, which leverage microphone arrays, enable precise spatial mapping of sound events, enhancing the contextual understanding of city spaces.

Together, these advancements in both machine learning algorithms and hardware technologies are opening new opportunities to capture the fine-grained interactions that occur in urban environments. We are now able to leverage large datasets representing interaction among urban actors under different scenarios and analyze common patterns of specific populations, such as users of assistive tools, to propose interventions that aim to make these spaces safer and more accessible for underrepresented groups. These developments, however, also bring challenges that will likely drive research in areas such as computer vision and human-computer interaction in the coming years.

Building on our work deploying multimodal street-level sensors for large-scale data collection, we explore how the combination of audio and video data collected at the street level can offer deeper insights into urban accessibility and human behavior. In the sections that follow, we identify the key opportunities and challenges ahead, emphasizing the unique affordances that street-level multimodal datasets provide over traditional approaches.

2. Related Work

Several efforts aim to collect multimedia urban data and use this data to reassess and build safer, more equitable, and more accessible urban spaces. Different research communities have studied and analyzed urban environments for decades using various approaches. However, large-scale urban multimedia datasets have only recently been made publicly available. Focusing on sound, SONYC [1] allows for a city-wide longitudinal understanding of human activity based on audio events. This dataset motivated several custom-made tools for efficiently exploring such modality, like Time Lattice [2] and Urban Rhapsody [3]. Moving forward, initiatives like StreetAware [4] and AIWaysion[1] combine several modalities, including video, audio, and LiDAR, to support experts interested in understanding fine-grained dynamics of pedestrian activity, including crossing patterns. It is in the context of the experience of building the StreetAware dataset that this piece reports opportunities and challenges to leverage video and audio data to explore urban dynamics.

3. Opportunities

In this section, we elaborate on the opportunities we envision for extracting useful information from recent audio-visual datasets, such as StreetAware. We start by indicating its advantages over other types of data and then show how it can be used to improve pedestrian safety under various scenarios.

3.1. Affordances of Audio and Video

Aerial versus Street-Level. As previous works present [5], aerial imagery can greatly support vehicular flow analysis. However, given its distance from objects of interest, we miss granular details that can enhance our comprehension of the scene. Pedestrian movement patterns, gait information, sidewalk materials, and street-level signal information represented in traffic lights are missed in this kind of data but are present in street-level imagery. This information is crucial to identify scenarios where pedestrians with disabilities are using different urban spaces. Enabling the automatic identification of classes related to pedestrians with disabilities (e.g., wheelchair and walker users) can support researchers in summarizing cases where these citizens are present and, therefore, understanding the common characteristics of these actors while interacting with various urban regions.

LiDAR versus RGB Video. LiDAR is a powerful technology that users laser imaging to quickly record data and convert build three-dimensional point clouds. This technology, however, still presents weaknesses for the work we want to accomplish in urban environments.

The dataset from Toronto-3D [6] is an example of comprehensive LiDAR data collection; it covers one kilometer of a road in Toronto, Canada and includes a dense point cloud with manually labelled segmentation of interesting features, such as road, building,and utility line.

The authors of Toronto-3D, however, have pointed out to know issues of the dataset involving moving cars. For our research purposes, the motion of vehicles is crucial to explore the interaction between motorized and non-motorized actors in cities. With multiview RGB video data, we can track moving vehicles and easily compute direction of mo-

[1] https://www.aiwaysion.com

tion, speed, and how close these vehicles are to other agents. These metrics are necessary for accurate time-to-collision calculations, which often serve as proxies to dangerous urban encounters.

Audio. Video and LIDAR data can be key to reconstructing digital representations of urban environments. They are, however, limited to reconstructing what is in their field-of-view. With audio data and audio detection models (such as YAMNet [7]), we can "expand" the field-of-view of these sensors. Events like an emergency vehicle approaching an intersection at high speed can benefit from audio detection, as the model would point out to these out-of-sight events to enrich data analyses that would otherwise be lost to limited field-of-view video data. This modality enables analysis aiming to understand the reaction time of different groups to out-of-sight events, which can support urban planners in creating tools to increase spatial awareness of groups of pedestrians with lower mobility levels. Furthermore, audio localization techniques can support the visualization of spatial locations where drivers more often need to use attention-grabbing strategies, such as car honks. These can highlight spaces where interventions, like reducing the speed limit or improving signaling, can reduce the probability of accidents.

3.2. Movement Analysis

Personal smart devices, such as smartwatches and smartphones, often measure people's movements throughout the day, reporting health metrics relevant to a person's well-being. These data range from number of daily steps averaged in a week to heart rate throughout sleep. The data, collected by personal devices, are private and belong to the individuals themselves (in section 4.3 we discuss challenges surrounding privacy).

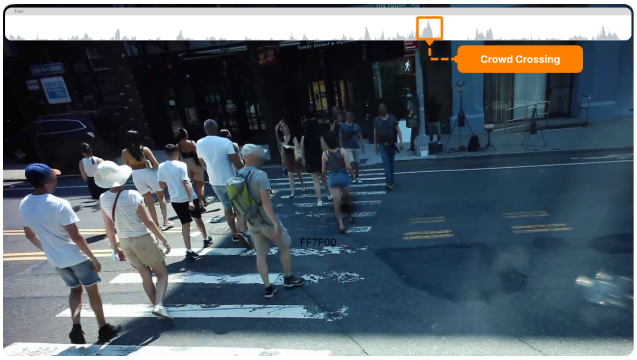

Figure 2. Distribution of pedestrian density over an entire video. Crowds are known to reduce the speed of pedestrians crossing the streets. Adapting traffic light timing based on pedestrian speed can be achieved by leveraging street-level RGB videos.

To extract similar movement data at large-scale in urban environments, we forego such personal devices and deploy our own cheap, multimodal sensors. Advancements in deep learning this past decade now allow for human pose prediction models [8–11]. With our RGB video data, we use such models to extract large-scale collections of two-dimensional poses of people on the streets. Using multiple of our multimodal sensors recording from different views, we can then reconstruct people and capture their movements in the three-dimensional space. We then extract metrics of people's movements patterns, such as speed of movement, direction, and gait. For gait, for instance, we can measure distance between consecutive steps and the double support (DS) time. DS time is the phase of the gait cycle where both feet are on the ground, allowing for greater stability and control of direction [12]. Increased DS and large DS variability have both been associated with higher fear of falls and risks of falling [13, 14]. Therefore, we find that measuring gait of people accurately and at large scale in urban environments allows for urban planners to prepare and adapt the current infrastructure to benefit people of all motor skills.

This type of granular gait data also opens the opportunity to explore how people move in different weathers, obstacles, and pavement materials. Project Sidewalk [15] has developed a crowd-sourced, scalable data collection of sidewalk accessibility information using street-view images. CitySurfaces [16] automatically detects, computes, and segments different pavement materials of sidewalk from street-level images. Our work can complement these two works with the extraction of granular gait pattern of pedestrians. By combining the pavement materials, accessibility issues of sidewalks, and the gait data collected by our multimodal, street-level sensors, experts can now study at large-scales the movement pattern of individuals with distinct mobility levels on, for instance, sidewalks with surface problems during rainstorms. These opportunities of data analyses highlight the role of urban planners and experts to better prepare our cities to accommodate different movement abilities in diverse (often adverse) environmental conditions and sidewalks with a range of accessibility issues.

3.3. Adaptive Signal Timing

Our proposal can open many pathways to interesting future developments. With the advancement of Artificial Intelligence and faster network connectivity, street-level multimodal sensors can be integrated into the traffic system.

Audio. Detecting audio events allows for a more dynamic and responsive traffic management, where, for instance, if the sensor hears an emergency vehicle siren approaching the intersection, it can safely adapt traffic and pedestrian crossings to allow for faster response time and reduce the chance of any unnecessary harm at the intersection.

RGB Video Data. With high-resolution cameras, we can monitor traffic light signals by extracting accurate traffic

light cycle timing information. This new feature permits in-depth analysis of traffic management systems, crossing patterns, and vehicular flow at intersections. Street-level multimodal sensors can, in the future, be integrated with traffic management systems to allow pedestrians, especially those with lower mobility, to wait less and to cross the intersection with enough time. Using the aforementioned techniques to reconstruct gait patterns and to identify assistive tools correlated with lower mobility levels, such an intelligent system could leverage this information as a proxy for the presence of low-mobility actors to adapt traffic light times.

4. Challenges

4.1. Expert in the Loop

Making these data useful for a diverse and non-technical community of domain experts is not trivial. Allowing these experts to directly interact with the data is an important step towards empowering them to make more informed decisions. However, extracting meaningful information from a mass of complex and large data is a difficult task. Therefore, designing intuitive visual analytics systems able to summarize and underline important details contained in the data is a must. These systems should, among other things, support cross-modal and interactive queries, requiring specialized data management infrastructure. Furthermore, tailored visualizations to enable the interpretation of different streams containing audio, visual, and derived metrics should be adequately tested against the target audience to assess their usability.

4.2. Optimal Deployment

We have, for now, a finite number of sensors which can record one hour uninterruptedly. This poses a significant bottleneck in our data acquisition. We have continuously worked to improve these sensors, including higher camera resolution, longer battery life, and more resilient to extreme weather. In order to obtain meaningful metrics of population with different abilities, the sensors need to be carefully placed around the city. We expect that by selecting neighborhoods with distinct socioeconomic conditions and selecting locations near schools, hospitals, and assisted living houses, our studies will more likely represent people with different crossing abilities. We also expect to highlight inequities in people's movement patterns around the city. Some of these might include longer crossing wait times for pedestrians, more honks and higher decibel levels of noise, and poorer infrastructure and accessibility (such as lack of appropriate signage and crosswalk paint) for under-resourced communities. We also will guide the deployment by the occurrence of traffic violations and accidents around the city. We aim to use publicly available data from *NYC Open Data* to identify such landmarks and study which areas have a greater risk of harm that should be prioritized for sensor deployment.

4.3. Privacy

Our work entails recording hours of human behavior in urban environments. Often, the metrics we are interested in exploring involve outliers and anomalies. These can range from pedestrians crossing while distracted by their cellphones to cars violating traffic signaling. We have taken the steps of anonymizing the video data by blurring faces and anonymizing the audio data by deleting recognizable speech. This might prove insufficient with the general population, as they can rightfully be skeptical of such data collection in public spaces and might behave differently due to the presence of sensors in their surroundings. Additionally, further study is needed to understand whether two-dimensional and three-dimensional poses can provide personally identifiable information. Thus, implementing a citizen science agenda to better understand the general population's perception of privacy-preserving concerns will be useful in guiding future deployments and communication strategies.

5. Conclusion

In this paper, we presented a set of opportunities the democratization of large-scale street-level multimodal data will enable over the next few years. Together with these, a set of challenges needs to be addressed to efficiently and responsibly support domain experts and policymakers in developing equitable and accessible regulations to support the safety and inclusion of every individual in the urban environment.

References

[1] Juan P. Bello, Claudio Silva, Oded Nov, R. Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. Sonyc: a system for monitoring, analyzing, and mitigating urban noise pollution. *Commun. ACM*, 62(2):68–77, January 2019. 2

[2] F. Miranda, M. Lage, H. Doraiswamy, C. Mydlarz, J. Salamon, J. Lockerman, Y. Freire, and C. Silva. Time lattice: A data structure for the interactive visual analysis of large time series. *Computer Graphics Forum (EuroVis '18)*, 37(3):13–22, 2018. 2

[3] Joao Rulff, Fabio Miranda, Maryam Hosseini, Marcos Lage, Mark Cartwright, Graham Dove, Juan Bello, and Claudio T. Silva. Urban rhapsody: Large-scale exploration of urban soundscapes. *Computer Graphics Forum*, 41(3):209–221, June 2022. 2

[4] Yurii Piadyk, Joao Rulff, Ethan Brewer, Maryam Hosseini, Kaan Ozbay, Murugan Sankaradas, Srimat Chakradhar, and Claudio Silva. Streetaware: A high-resolution synchronized multimodal urban scene dataset. *Sensors*, 23(7):3710, April 2023. 2

[5] Eugen Valentin Butilă and Răzvan Gabriel Boboc. Urban traffic monitoring and analysis using unmanned aerial vehi-

cles (uavs): A systematic literature review. *Remote Sensing*, 14(3), 2022. 2

[6] Weikai Tan, Nannan Qin, Lingfei Ma, Ying Li, Jing Du, Guorong Cai, Ke Yang, and Jonathan Li. Toronto-3D: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 202–203, 2020. 2

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 3

[8] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. In *CVPR*, 2021. 3

[9] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, 2019. 3

[10] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023. 3

[12] Anup Nandy, Saikat Chakraborty, Jayeeta Chakraborty, and Gentiane Venture. *Modern Methods for Affordable Clinical Gait Analysis*. Academic Press, 2021. 3

[13] Daniel S. Williams and Anne E. Martin. Gait modification when decreasing double support percentage. *Journal of Biomechanics*, 92:76–83, July 2019. 3

[14] Jeffrey M. Hausdorff, Dean A. Rios, and Helen K. Edelberg. Gait variability and fall risk in community-living older adults: A 1-year prospective study. *Archives of Physical Medicine and Rehabilitation*, 82(8):1050–1056, 2001. 3

[15] Manaswi Saha, Mikey Saugstad, Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, and Jon E. Froehlich. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. 3

[16] Maryam Hosseini, Fabio Miranda, Jianzhe Lin, and Claudio T. Silva. Citysurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society*, 79:103630, April 2022. 3