What TripAdvisor Can't Tell:
Crowdsourcing Urban Impressions for Whole Cities

Daniel Gatica-Perez[1], Salvador Ruiz Correa[2], Darshan Santani[3]

Summary

In the context of a number of emerging opportunities for citizen participation in Latin America, we present an overview of the SenseCityVity project developed in Mexico. The project goals are the design, implementation, and validation of a mobile crowdsourcing framework involving youth taking photographs with smartphones and providing online impressions about the city as captured in geo-referenced photos. One key aim is to co-design experiences with youth that help make visible the specific urban issues that matter to them, and which are seldom present in other sources of social media data. The project activities include the mapping of issues including accessibility, safety, or trash management; the use of statistical data analysis methods to assess the reliability of the provided impressions and the associated ability of the observers; and a reflection process through which proposals to work on such issues can potentially be generated.

Introduction

Geo-localized social media like Foursquare and Yelp, in which people talk about urban places, and more traditional online media like TripAdvisor, are at their core *crowdsourced mechanisms to document the city*. In all these services, people voluntarily contribute information in the form of check-ins, comments, and reviews about the places they visit. This information reveals personal states of mind, taste, sentiment, and opinion. In principle, this information is contributed for free, although some of the above services have also explored ways of incentivizing social participation using various mechanisms, monetary or otherwise.

The above services ultimately respond to an economic agenda, which can benefit patrons and business owners. As such, geo-localized social media have inherent dynamics that promote participation on urban areas of commercial and touristic use, and often high socio-economic status. When collected and aggregated with the use of machine learning and data mining methods, this data can provide a detailed picture of what these *very specific* urban areas are about. Due to this, they represent a valuable resource to study digital cities at scale.

Needless to say, however, cities are more than their commercial or touristic spots. We posit that urban crowdsourcing can be used to document other places and aspects of cities, including what tourists and higher-income locals do not often see or talk about. This includes entire areas of cities that are not popular among tourists or wealthier locals (e.g. areas that might be in disarray or inaccessible), and that, because of this very reason, remain to a large extent invisible in social media. This is especially important for many cities in Latin America, where international tourists and booming youth local populations share the same urban space, physically and online, both being the most active contributors of geo-localized social services.

In this paper, we present an overview of SenseCityVity (**http://www.idiap.ch/project/sensecityvity/**), an ongoing interdisciplinary project in Mexico where we integrate research in computer science, social psychology, and urban studies, with the goals of designing, implementing, and validating crowdsourcing experiments on urban impression formation from photographs taken by local youth with mobile phones. One of our aims is that youth participation in mobile crowdsourcing helps render visible in social media a number of urban issues that matter to them. This includes the mapping and documentation of issues in the city that need to be improved regarding security, accessibility, or trash management; and a reflection process through which discussions and proposals to better address such issues can emerge.

[1] Idiap Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland.
[2] Centro Nacional de Supercomputo (CNS-IPICYT), San Luis Potosi, Mexico.
[3] Idiap Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland.

The article is organized as follows. In Section 2, we summarize current strategies in social media to characterize cities. In Section 3, we introduce the SenseCityVity project, and briefly discuss the project main objectives and findings. Section 4 provides some concluding remarks.

2. A biased tale: social media data to characterize digital cities

Three sources of urban information currently used to characterize complex phenomena in digital cities are mobility data, phone data, and social media data. First, mobility data is growing due to the number of sensors monitoring human travel, including public transportation cards and sensors on roads. In urban computing, urban mobility has been studied using taxi traces, automated fare collection metro cards, and shared bicycle systems (Froehlich, 2009). These sources of data are comprehensive with respect to the urban areas that can be covered; on the other hand, they typically have limitations of temporal resolution (e.g. only check-ins are commonly available in metro cards) and of spatial resolution (e.g. only beginning and end points are available in shared bike data). While this is enough to identify temporal patterns around transportation hubs, it limits the study of fine-grain trajectories.

Second, there has also been a surge of interest in using mobile phone operator data (call detail records or CDRs) to characterize urban mobility, in which traces can be extracted from cell tower connectivity when people use their mobile devices (Becker, 2013). This source of information, while attractive in terms of scale, has issues related to privacy, as well as limitations in terms of spatial resolution (for instance, it cannot be effectively used to characterize fine-grain location within public spaces). Urban mobility has also been captured using smartphone sensors. A recent example of urban scale data collected from smartphones is the work in (Laurila, 2013), which collected 24/7 smartphone data (location, motion, communication) from 180 volunteers for over a year of time in French-speaking Swiss cities and towns. While this approach can generate a variety of high-quality data for research, scale remains as one of its limitations, as it requires its widespread adoption by citizens, who justifiably have concerns, ranging from short-term (battery life) and long-standing ones (privacy).

Finally, mobile social media is an alternative to collect large-scale urban data about mobility and public spaces, exemplified by Foursquare check-ins, geo-localized Tweets, and geo-localized images (Cho 2011). As discussed in the introduction, mobile social media sites are especially active in popular urban spots (outdoor hotspots for young people, attractions for locals and tourists, major events, and restaurants and nightlife spots). Geo-social data has clear advantages in terms of built-in mechanisms for social acceptability and engagement, fine spatial resolution, and links between the physical and the online worlds, providing additional online sources of text and multimedia. On the other hand, this data source suffers from temporal sparsity as check-ins and photo-taking occur sparsely over time for any given user.

As an example, Figure 1 shows snapshots from Foursquare and Yelp from Guanajuato City, Mexico. Each of the sites provides photos, comments, and interaction about popular places in the city. Guanajuato City (pop. 170,000) is the capital city of Guanajuato state in Central Mexico. The city is well known for its art scene and tourism (notably, the City hosts each year during the month of October the International Cervantino Festival, the largest art festival in Latin America, attracting visitors in the hundreds of thousands), yet it also has a variety of socio-urban problems. The city is located in a little valley with narrow and winding streets, as well as pedestrian alleys on the mountain sides, leading to houses with no car access and that correspond to urban sprawl. The historical center, dating from the Colonial period, has a multitude of plazas, temples, theatres, and government constructions. Furthermore, pedestrian alleys are another key urban feature of the city. The city reflects a common situation in cities in Latin America, where a combination of historical downtown areas, urban sprawl, increasing populations, and large socio-economic disparities all co-exist.

Foursquare and Yelp users are documenting Guanajuato's popular places. This includes the main landmarks (see references to *Centro Historico* and *Jardin de la Union* in both sites in Fig. 1) as well as restaurants, bars, and clubs (see references to *La Vie en Rose* and *La Clave Azul*). Furthermore, for some of these places, users have contributed hundreds of photos and thousands of visits (see Figure 2), often without realizing the implications that their individual actions represent at the collective level. In contrast, however, for much of the rest of the city, none of these sites provide any information. This can include places of the same kind (restaurants, bars, etc.) that are not popular among users of social media but that nevertheless could be

popular among other segments of the population in Guanajuato, as well as a multitude of non-touristic venues, anodyne streets, hidden urban gems, and places that are regarded by the locals as dangerous, dirty, or in disarray. Clearly, this is the result of socio-demographic biases that exist in social media, where not everyone is represented, not every place is represented, and not every activity is represented; these sources of bias are often not accounted for in urban computing research, despite the inevitably partial views that can be obtained from social data (Tufekci, 2014).

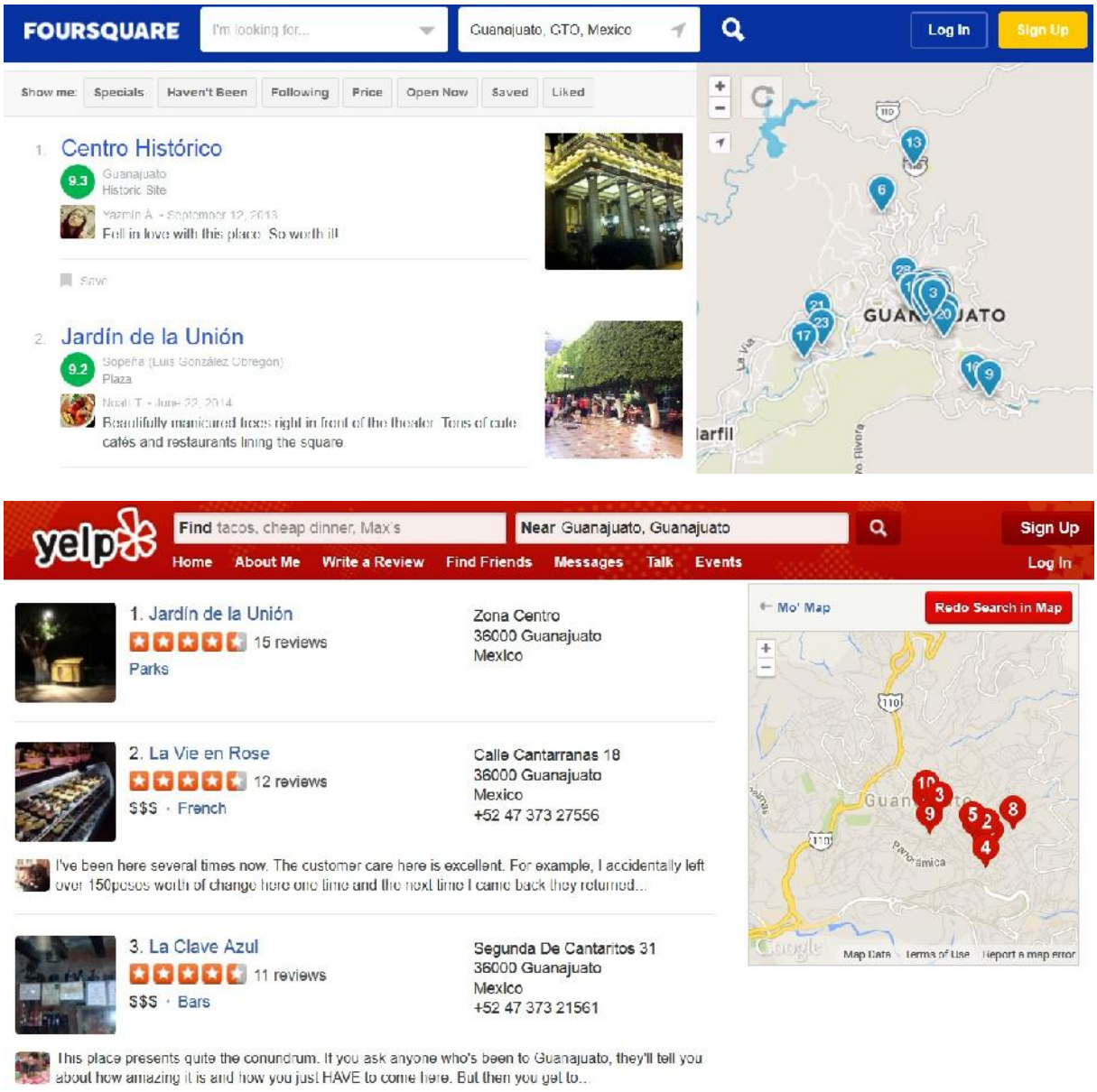


Figure 1. Snapshots from Foursquare and Yelp for Guanajuato City, Mexico (accessed May 2015).

In summary, much of the existing research in social and urban computing has studied the digital city using single sources of data, given the inherent difficulties associated to sensing and access, and the potentially daunting scales. However, it is clear that an integrative approach, where multiple sources of urban information are combined, would allow to paint more accurate pictures of cities, and to reduce the biases that each of the individual sources contain. A recent example in this direction is (Antonelli, 2014), in which Telecom Italia released for research two months of proprietary phone operator data for two Italian cities in

combination with Twitter data, weather data, etc. As an alternative direction, we propose to combine social media data and citizen-contributed mobile data to capture, document, and reflect on issues of relevance to urban dwellers, thus reducing the inherent biases in other digital city data sources. We present this approach in the next Section.

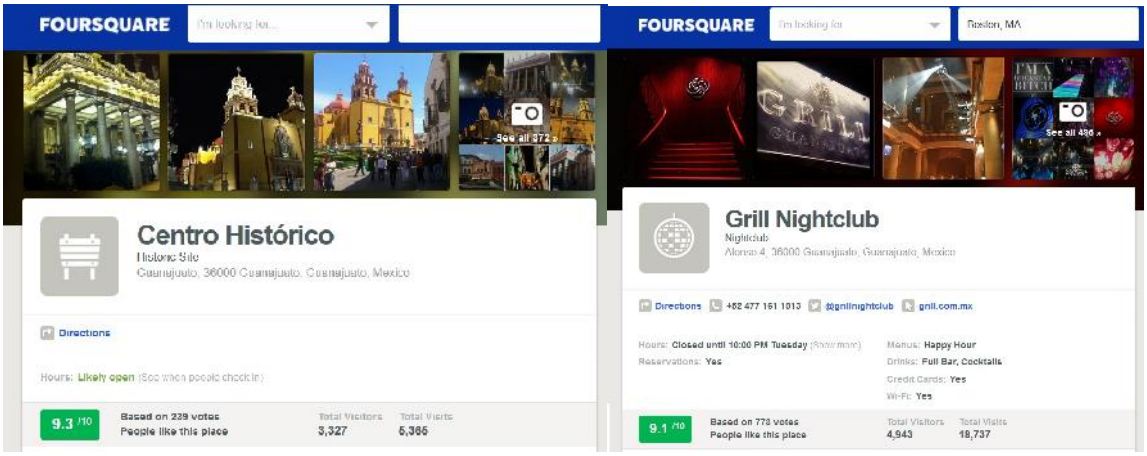


Figure 2. Foursquare snapshots of *Centro Historico* and *Grill NightClub*, Guanajuato (accessed May 2015).

3. The SenseCityVity project

SenseCityVity was a collaboration project between Idiap-EPFL in Switzerland, and the National Supercomputing Center in Mexico (CNS-IPICYT), which was supported during 2104 by the Cooperation and Development Center at EPFL. The aim of the project was to investigate the feasibility of design and rapid deployment of participatory mobile sensing and crowdsourcing technologies in Guanajuato City, in order to document and address specific socio-urban concerns and perceptions of youth populations. Dimensions of place included accessibility, safety, cleanliness, preservation, beauty, and interest. A number of computational analyses conducted on *SenseCityVity* data led to an improved understanding on how young people grasp their environment according to the above dimensions. Statistical, crowdsourcing, and ethnographic analyses allowed to quantitatively determine what categories of urban concerns are noticed, which ones are perceived as more prevalent or pressing, which ones are perceived as having significant socio-economic impact, and which ones go unnoticed by the population participating in the study.

3.1 Goals

Major challenges in attaining sustainable development, strong economic growth, and social wellbeing in developing countries are closely related to the state of the urban environment in cities, neighbourhoods and communities. For this reason, the use of methodologies leading to an improved understanding of socio-urban problems and citizen concerns has significant value. SenseCityVity took these matters upon by addressing specific urban issues jointly with a population of high-school students from Guanajuato City, through the use of participatory mobile sensing and crowdsourcing technologies. We had three specific goals:

(1) Design a framework of interdisciplinary nature that included a research team and actors, and that was the basis for the whole project work, starting with the definition of the urban concerns.
(2) Collect a dataset of geo-localized images and videos collected through mobile crowdsourcing and depicting urban concerns in Guanajuato City, including photos of places and videos where students discuss.
(3) Analyze the collected data from the perspective of how local students perceive the urban space captured in geo-localized photos and videos with machine learning, crowdsourcing, and ethnographic techniques.

Each of these goals and a summary of the corresponding results are described in the next subsections.

3.2 Framework

A population of student volunteers who participated in the co-design of the sensing and crowdsourcing phases were recruited from the *Centro de Estudios Cientificos y Tecnologicos campus Guanajuato* (CECYTE), a technical high school in Guanajuato City that has a population of 600 students. The research team worked with the students to define the specific urban concerns to be addressed by the population, the type of relevant data that would be collected, and the specific use of the resulting crowdsourced resources. A key element of our project was that, instead of a top-down approach where target concerns were pre-defined, the participants defined them by themselves. These concerns were clearly articulated during a series of workshops with researchers, in which topics such as urbanism, participatory sensing, data privacy, and the use of mobile technology to explore the urban environment were addressed. The creative use of the collected data also led to a Film Festival in which teams of student volunteers (supervised by the research team) presented several short video documentaries that summarized their views of the city sites explored during the mobile sensing experience.

3.3 Data collection

The data collection in the project took the form of an Urban Data Challenge (UDC). During the UDC, ten teams of ten members each walked across the city in order to capture images and videos of urban sites. In most cases, teams started their work in the early morning, mostly on weekends, and followed traditional routes to cover various areas of the city, mainly in the downtown and surrounding areas. Figure 3 shows typical trajectories followed by the teams in the downtown area. In the figure, each trajectory is identified by circles of a specific colour. As Guanajuato City is surrounded by hills, some of the mobility patterns indicate that teams travelled some distance from an initial point and move toward a specific area of interest, which in many cases is located on the top of a hill, so there is not a large variation in latitude-longitude coordinates. Mobility patterns also include those in which teams travelled a long distance from an initial point starting in the downtown area through the city's main streets, which are either almost straight lines or loops surrounding the historic downtown area (not shown in the figure).

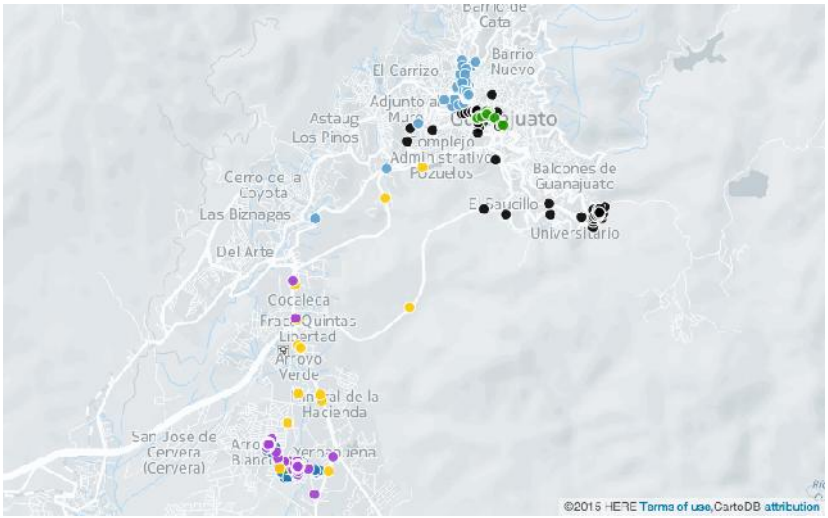


Figure 3. Typical mobility patterns followed by SenseCityVity teams during the data collection process in Guanajuato City downtown area (Urban Data Challenge). Each trajectory is identified by a specific colour.

The effort by the students and the research team resulted in a rich data set consisting of geo-located photos, socio-urban video clips, GPS locations, video interviews, geo-localized tweets in Guanajuato City, crowdsourcing judgments (corresponding to annotations of a primary image dataset, in which local students expressed their impressions about urban spaces), and video documentaries created by the participants themselves. The exact data amounts are shown in Table 1.

Datatype	Description
Geo-localized photos	7,000
Socio-urban video clips	108
Video interviews	186
GPS locations	5,298
Geo-localized tweets	18,328
Facebook posts and total post reach	72 and 9,637
Crowdsourcing judgements	9,027
Video documentaries	13

Table 1. SenseCityVity data set. The exact number of items for each data type is shown in the right column.

3.4 Data analysis

Mapping urban concerns. We first computed a heat map of geo-localized images to identify areas where the photos were taken, thus corresponding to areas of concern. The map was computed using kernel density estimation techniques for spatial data applied on the GPS locations of the image data set. The areas of concern obtained during the mobile crowdsourcing experiment are shown in the heat map of Figure 4 (left). The sites shown in red in the Figure correspond to traditional neighbourhoods, plazas, city alleys, and centric avenues.

The information provided by these heat map contrasts to what is obtained by aggregating a set of generic geo-localized tweets for the same geographical area, collected with the public API and corresponding to a three-month period in 2014, as depicted in Figure 4 (right). As expected, most tweets occur in the touristic zone of the city and the University of Guanajuato main campus. As discussed in the introduction, mobile social media data from Twitter can be sparse in certain areas of the city where the student teams detected urban concerns. This shows the relevance of our approach.

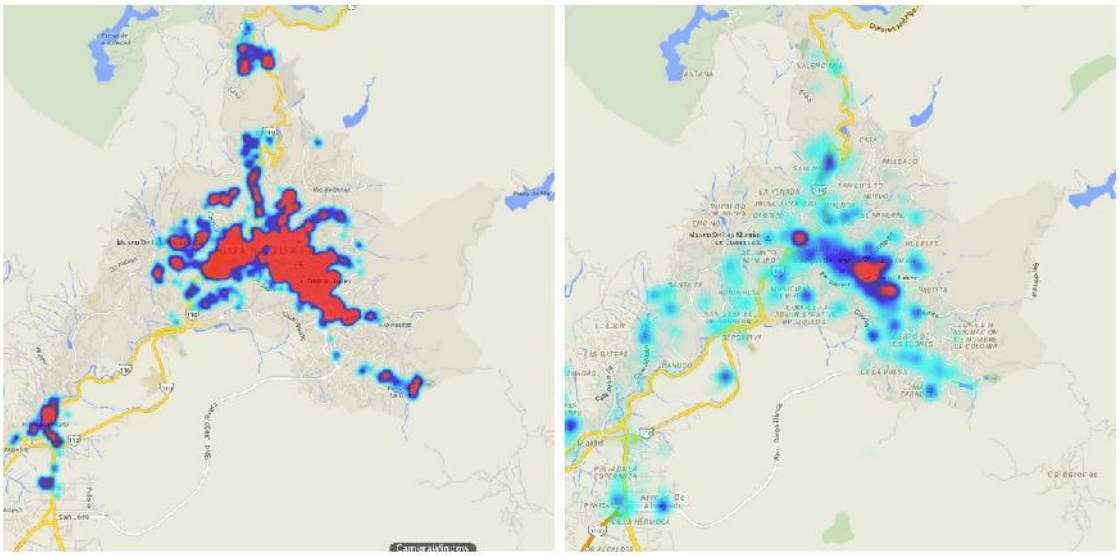


Figure 4. Left: Heat map computed from GPS locations of crowdsourced images. The map illustrates the areas of urban concerns detected during the data collection process. Right: Heat map of geo-localized tweets (generated independently of our project.)

Analysis of video interviews. We conducted an ethnographic study of the video interviews addressing youth urban concerns. This analysis led to a list of urban problems listed in Table 2. The number of individuals who mentioned a specific problem is shown in the plot of Figure 5. Problems are divided into three main categories: city image, city infrastructure, and city life. Each of the bars shown in the plot of Figure 5 are coloured according to the category they belong, in green, blue, and orange, respectively.

Garbage in streets and alleys and non-artistic graffiti are two of the urban problems of most concern for the interviewed youth. Many of them are preoccupied about the image of a touristic city in view of these pervasive problems. Several peopled agreed with the view that "[some] streets are very ugly and full of trash; people do not have education and do not realize that trash and graffiti damage the image of a touristic city," as stated by one of our participants.

Observed urban problems	
City image	1. Garbage in streets and alleys. 2. Non-artistic graffiti. 3. Dirty streets. 4. Pollution. 5. Street dogs. 7. Street odors. 6. Pests (rats and cockroachs).
Infrastructure	8. Seriously worn city infrastructure (alleys, streets, houses, and buildings). 9. Insufficient garbage containers; inefficient garbage collection. 10. Insufficient street lightning in alleys and some streets. 11. Garbage in sewers leading to street floodings. 12. Inaccessibility issues in street alleys. 13. Inadequate high power electric wiring over posts and houses. 14. Inadequate sewerage in city alleys.
City life	15. Crime, vandalism, and gangs. 16. Feeling of insecurity. 17. Insufficient police surveillance. 18. Insufficient public transportation. 19. Inefficient public transportation. 20. Alcoholism and drugs in streets and alleys. 21. Indigents. 22. Traffic. 23. Violence in streets (gang fights).

Table 2. Urban problems detected by Urban Data Challenge participants.

The lack of enough garbage containers and the location of those that are available are also main concern to the inhabitants of Guanajuato City, as confirmed by the results in Figure 5. More containers and a more efficient way of managing waste could significantly improve the image of the city, particularly in city alleys: "the people in charge of collecting garbage in the city are not able to do it on time, and the garbage containers often overflow polluting the environment," as stated in one of the interviews.

The participating youth also observed that many streets, buildings, and houses are worn out and require urgent repairs. Although the downtown area is in general in good shape, most places of the city around downtown, particularly historic city alleys, are severely deteriorated, as stated in one interview: "the quality of housing, health, and education of native people or people coming from outside, who live in the alleys even close to the downtown area is very poor; these adverse circumstances lead to violence."

Youth also complained about a generalized sense of insecurity, vandalism, and gang activity in many areas of the city. They feel that prevention of addictions and increased police surveillance are an urgent need. One participant stated: "The problem here [in the city] is that there is a lot of insecurity, in the alleys, outside the downtown area. At night, there are people drinking and smoking marihuana in the street alleys. We used to play by Hidalgo market every day. But now we are limited because of insecurity, the *cholos* [gangs]; things can be complicated and we are limited to go outside."

A problem that many people in Guanajuato endure every day is the lack of adequate urban transportation, particularly for those traveling across the city to go to school or work. This was confirmed by the results in Figure 5, and illustrated by this statement: "Transport does not come in time or is very scarce, or things like that, which affect us to get to school; this problem affects me directly and I think it affects the majority of us [youth]". Overall, the participating population in our study was concerned about pressing urban problems in Guanajuato City, such as garbage in the streets, non-artistic graffiti, crime, insecurity and worn infrastructure.

However, as Figure 5 also reveals, the population was not as aware with regard to other problems that affect the city significantly. The SenseCityVity team working in the field also detected important urban problems, which are not listed in Table 2. For instance, the city lacks infrastructure for the disabled, has increased levels noise, visual pollution, and air pollution in streets and city's tunnels.

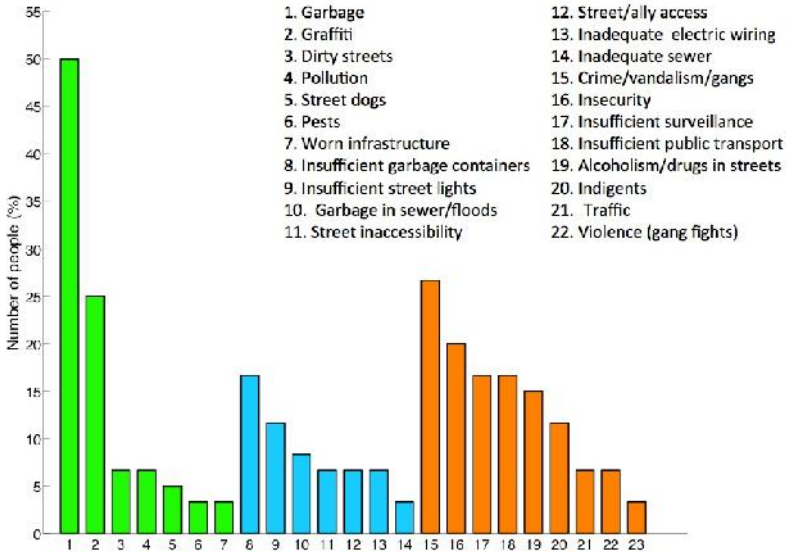
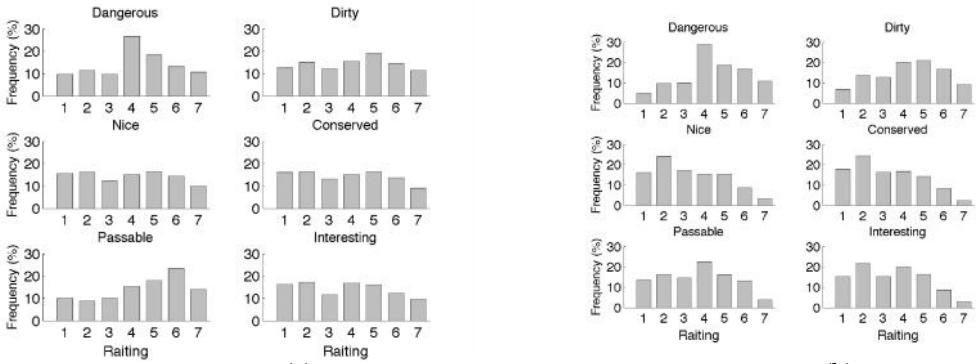


Figure 5. Urban problems detected by Urban Data Challenge participants. Problems are divided in three categories: city image (green), city infrastructure (blue), and city life (orange).

Analysis of crowdsourced impressions of urban places. We summarize some of the findings related to the analysis of crowdsourced local impressions provided by the SenseCityVity participants and the statistical modelling of annotator responses. A more detailed description can be found in (Ruiz-Correa, 2014). We conducted a crowdsourcing study to gather 9027 impressions of urban sites involving over 150 high-school students, using standard scales in social psychology for 102 images, which allowed us to use standard inter-rater reliability and response analyses. Six dimensions were annotated on a Likert scale (dangerous, dirty, nice, conserved, passable, and interesting.) The histograms of mean annotator ratings are shown in Figure 6. A statistical analysis indicates that none of the histograms correspond to a uniform distribution (Chi square test, $p < .00001$). We found that the crowdsourced annotations have good levels of reliability (Shrout, 1979), which points towards the feasibility of collecting this type of perceptions from local populations (Santani, 2015). Furthermore, we successfully fitted a latent variable model commonly used in psychometric studies that allowed for the characterization of an annotator's ability to make judgments as well as the difficulty in rating images depicting urban sites (Rasch, 1980).



(a) (b)

Figure 6. Mean ratings histograms for urban descriptors associated with a set of SenseCityVity images and the student annotations. Histograms in (a) correspond to images of city streets and plazas, whereas the ones in (b) correspond to images of city alleys. A seven-point Likert scale ratings were obtained for six urban descriptors (dangerous, dirty, nice, conserved, passable, and interesting).

4. Conclusion

Cities are more than their commercial or touristic spots, and the digital expression of cities is more than online advertising and personalized recommendations. We have shown that mobile crowdsourcing in cities can be used to contribute to this diversity, documenting urban spaces which high-income locals and tourists do not often see or want to talk about. This is especially relevant for cities in Latin America, where economic disparity translates into digital invisibility. The SenseCityVity project has demonstrated, in a specific local context, the possibility of designing and deploying a platform for mobile data collection and analysis that fosters young citizen participation; the use of these technologies to facilitate the documentation of a number of socio-urban citizen concerns; and the potential of crowdsourcing and data mining to contribute to the understanding of the urban space as perceived by a population. In particular, the possibility of combining citizen-contributed data with machine learning and statistics can help reveal what existing urban problems are perceived as more prevalent or go unnoticed, how these problems are geographically stratified, and what types of general impressions a population has about its own urban space. Further analyses of these issues are part of our future work.

Acknowledgments

We thank our colleagues from the SenseCityVity team for their contribution to the project: Itzia Ruiz Correa (SenseCityVity.Media), Carlo Olmos Castillo (SenseCityVity.Art), Brisa Carmina Sandoval Mexicano (SenseCityVity.People), Beatriz Ramirez Salazar (SenseCityVity.Support), and Juan Luis Salazar Villanueva (SenseCityVity.Apps). We also thank the student population at CECYTE for their enthusiastic participation and key contributions, and the authorities of the CECYTE for the support provided to our study.

References

Fabrizio Antonelli, Matteo Azzi, Marco Balduini, Paolo Ciuccarelli, Emanuele Della Valle, and Roberto Larcher, City sensing: visualising mobile and social data about a city scale event, in Proc. International Working Conference on Advanced Visual Interfaces (AVI), May 2014.

Richard Becker, Ramon Caceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky, Human mobility characterization from cellular network data, Communications of ACM, 56(1):74-82, January 2013.

Eunjoon Cho, Seth A. Myers, and Jure Leskovec, Friendship and mobility: User movement in location-based social networks, in Proc. SIGKDD, pp. 1082-1090, 2011.

Jon Froehlich, Joachim Neumann, and Nuria Oliver, Sensing and predicting the pulse of the city through shared bicycling, in Proc. IJCAI, pp. 1420-1426, 2009.

Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh Minh Tri Do, Olivier Dousse, Julien Eberle, and Markus Miettinen, From big smartphone data to worldwide research: The mobile data challenge, Pervasive and Mobile Computing, 9(6):752-771, 2013.

Georg Rasch, Probabilistic models for some intelligence and attainment tests, The University of Chicago Press, 1980.

Salvador Ruiz-Correa, Darshan Santani, and Daniel Gatica-Perez, The Young and the City: Crowdsourcing Urban Awareness in a Developing Country, in Proc. Int. Conf. on Internet of Things in Urban Space, Rome, Oct. 2014.

Darshan Santani and Daniel Gatica-Perez, Loud and Trendy: Crowdsourcing Impressions of Social Ambiance in Popular Indoor Urban Places, in Proc. ACM Int. Conf. on Multimedia, Brisbane, Oct. 2015.

Patrick E. Shrout and Joseph L. Fleiss, Intraclass correlations: uses in assessing rater reliability, Psychological Bulletin, 86(2):420, 1979.

Zeynep Tufekci, Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls, in Proc. AAAI ICWSM, Ann Arbor, Jun. 2014.

Authors biography

Daniel Gatica-Perez is Head of the Social Computing Group at Idiap and Professeur Titulaire at EPFL in Switzerland. He works on methods that integrate social media, sensor data, machine learning, and social sciences to understand human behavior and to create applications for social good. His current work includes the study of urban trends using mobile social media and crowdsourcing, the analysis of conversational behavior in social video, and the understanding of emerging social phenomena in face-to-face interaction.

Salvador Ruiz Correa is Adjoint Professor and Research Scientist at the Instituto Potosino de Investigación Científica y Tecnológica (CNS-IPICYT) in Mexico, and codirects Ce Mobili (Center for Mobile Life), an initiative that studies urban environments with mobile and social technologies and designs applications that empower citizens and their communities in Latin America. He has a PhD in Electrical Engineering from the University of Washington, USA.

Darshan Santani is a PhD candidate at Idiap and EPFL in Switzerland. His research interests lie broadly at the intersection of computer science and social science, in particular the application of large-scale mobile data mining to various social and urban contexts. He received a M.S. on Management, Technology, and Economics from ETH Zurich, Switzerland, and a B.E. in Computer Science and Engineering from Manipal Institute of Technology, India.