Looking South: Learning Urban Perception in Developing Cities

DARSHAN SANTANI, Idiap Research Institute, Ecole Polytechnique Federale de Lausanne
SALVADOR RUIZ-CORREA, IPICYT, Mexico
DANIEL GATICA-PEREZ, Idiap Research Institute, Ecole Polytechnique Federale de Lausanne

Mobile and social technologies are providing new opportunities to document, characterize, and gather impressions of urban environments. In this paper, we present a study that examines urban perceptions of three cities in central Mexico, which integrates a mobile crowdsourcing framework to collect geo-localized images of urban environments by a local youth community, an online crowdsourcing platform to gather impressions of urban environments along twelve physical and psychological dimensions, and a deep learning framework to automatically infer human impressions of outdoor urban scenes. Our study resulted in a collection of 7,000 geo-localized images containing outdoor scenes and views of each city's built environment, including touristic, historical, and residential neighborhoods; and 144,000 individual judgments from Amazon Mechanical Turk. Statistical analyses show that outdoor environments can be assessed in terms of inter-rater agreement for most of the urban dimensions by the observers of crowdsourced images. Furthermore, we proposed a methodology to automatically infer human perceptions of outdoor scenes, using a variety of low-level image features and generic deep learning (CNN) features. We found that CNN features consistently outperformed all the individual low-level image features for all the studied urban dimensions. We obtained a maximum R^2 of 0.49 using CNN features; for 9 out of 12 labels, the obtained R^2 values exceeded 0.44.

Additional Key Words and Phrases: Mobile Crowdsourcing; Urban Perception; Outdoor Places; Deep Learning; Collective Action; ICTD; Mexico

ACM Reference format:
Darshan Santani, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2018. Looking South: Learning Urban Perception in Developing Cities. *ACM Trans. Soc. Comput.* 1, 1, Article 1 (January 2018), 23 pages.
https://doi.org/10.1145/3216367

1 INTRODUCTION

Community awareness and action on urban problems are long-standing practices in developing countries [20]. The ability to reflect and act upon concerns defined by a community's interests and values for its own benefit takes on special relevance in Latin America due to the local governments' inability to realize the full potential of both human and economic resources and a historically slow (when not absent) response by the authorities. Civic engagement and action with the local environment have educational, social, and economic aspects [14].

In this context, mobile and social technologies are providing new opportunities to document, characterize, map, and ultimately address urban problems in developing cities. Mobile crowdsourcing efforts for urban mapping and surveying conducted by citizens equipped with mobile phones

2469-7818/2018/1-ART1 $15.00
https://doi.org/10.1145/3216367

(who generate reports, take pictures, or create maps) are emerging, often concentrated in informal settlements and other problematic regions [1, 4, 45, 59]. Other recent approaches are studying cities in the developed world, and use online crowdsourcing platforms to establish the feasibility of obtaining estimates of urban impressions of physical and psychological constructs like safety, beauty, and quietness, elicited by images of the city taken at street level [50]. The possibility of obtaining crowdsourced perceptions of the socio-urban image of a city by non-residents is valuable for developing cities, especially when the cities have large flows of visitors (tourists, students or business people), as it could help to understand the choices that non-locals make regarding the use of the public space, or to identify misconceptions due to the lack of local context.

In research examining urban impressions in developed cities using online crowdsourcing, judgments have been elicited using images obtained from Google Street View (GSV) [23, 42, 50]. Even though GSV provides a scalable and automated way to collect images, it suffers from two limitations. First, the GSV image database is not exhaustive in spatial coverage in developing countries due to accessibility and safety issues. For instance, due to the way Google collects street views (via cameras mounted on top of a vehicle), GSV does not always contains images of narrow streets and winding alleys. In a previous study, we found that 58% of our images were either unavailable or erroneous in the GSV database for a mid-size touristic city in Mexico [45]. Second, GSV images fail to capture the temporal aspects of a city: only static views are available, and it can take years before images are updated. This does not facilitate studying the effect of time of the day in the perception of the urban environment, which is a key aspect as discussed in urban studies literature [31, 39]. In contrast, mobile crowdsourcing represents an opportunistic, just-in-time way of documenting urban changes over time.

In addition to collecting data, it is necessary to develop computational methods to automatically classify the perceptual attributes of outdoor environments. Gathering human judgments of such environments across multiple cities and subjective dimensions (e.g., dangerous, dirty, accessible, etc.) is both labor intensive and costly. However, recent advancements in deep learning have opened opportunities to use these techniques to automatically infer human judgments on subjective attributes, for instance urban perception labels as we study in this paper. Recent studies have shown that visual cues extracted using deep learning consistently outperform low-level images features, including color histogram in RGB space (Color), histogram of oriented gradients (HOG), dominant spatial structure of a scene (GIST), spatial arrangement of color and intensities (LBP), for a variety of computer vision tasks including inference of subjective properties [33, 38, 53]. In other words, generic deep learning features can potentially provide discriminative cues to perform automatic visual surveying of outdoor scenes.

In this paper, we present a study on urban perception of three cities in a developing country, which integrates: (1) mobile crowdsourcing involving a local youth community to collect first-person perspective images depicting urban issues that are defined by the community itself, (2) online crowdsourcing using Amazon Mechanical Turk (MTurk), where US crowdworkers contribute their impressions on photos of the urban environment along twelve physical and psychological dimensions of a place, and (3) a deep learning based framework to automatically infer human judgments of urban perception using visual cues extracted from crowdsourced images. We study Mexico as a representative case study of developing countries in Latin America, which has a large youth population (about 35% of the 120 million population are under the age of 18 [6]) and cities of different characteristics (agricultural, touristic, industrial), with a variety of urban issues. We address the following research questions in this paper:

RQ1: In the context of urban images collected in developing cities, what is the level of agreement among online crowd-workers with respect to the perception of physical and psychological dimensions of outdoor scenes? Are some of these perceptions different than the ones about developed cities?

RQ2: Using low-level image descriptors and deep learning models trained from data not originated from developing cities, can these crowdsourced perceptions of urban perception be automatically inferred?

To address these research questions, we make the following contributions:

(1) A mobile crowdsourcing framework involving over 70 local students that resulted in a data set of 7,000 geo-localized images collected in three cities in Guanajuato state of Mexico (Guanajuato, Leon, and Silao), each one characterized by distinct geography, economic activity, and population (Section 3). The data set contains outdoor scenes and views of each city's built environment, including touristic, historical, and business sites, residential neighborhoods, and areas with narrow streets and winding alleys. We are making the dataset (images and annotations) publicly available for research.

(2) An online crowdsourcing study on MTurk to gather impressions of crowd-workers along 12 physical and psychological labels including *dangerous*, *dirty*, *interesting*, *happy*, *polluted*, etc., based on 1,200 images (400 images per city) (Section 4). The studied dimensions include and extend those studied in recent literature. Statistical analyses on 144,000 individual judgments show that the outdoor scenes in the studied cities, when assessed by online crowd-workers, reach an agreement (measured by intraclass correlation, ICC) between 0.64 and 0.79, with 11 of the 12 dimensions reaching ICC agreement above 0.7 (Section 5). Furthermore, when comparing our data with an independent dataset collected in developed cities, we observe differences in perception with respect to a few dimensions in term of descriptive statistics, yet a principal component analysis shows a consistent structure of some of the dimensions across datasets, thus providing empirical support for a circumplex model of affect for physical environments (Section 5).

(3) Using the collected data, we proposed a methodology to automatically infer human perceptions of outdoor scenes, using a variety of low-level image features (including color histogram, texture, structural features, histogram of oriented gradients), and generic deep learning (CNN) features (Section 6). To extract CNN features, we used deep learning models pre-trained on data not originated from developing cities. CNN features consistently outperformed all the individual low-level image features for the 12 studied dimensions. We obtained a maximum R^2 of 0.49 using CNN features; for 9 out of 12 labels, the obtained R^2 values exceeded 0.44.

Overall, our current work engages youth communities as actors of social change and contributes towards understanding socio-urban problems in cities through the use of collective action, participatory sensing, and crowdsourcing technologies, and adds to the growing literature on automatic recognition of subjective attributes in urban scenes.

2 RELATED WORK

2.1 Systems for Reporting Urban Issues

There are various existing systems that allow citizens to report urban issues. One in the developed world is FixMyStreet (FMS) [3], launched in the UK in 2007 [29] and later implemented in other countries (mainly in Europe) with varying degrees of success. FMS allows people to share and map text reports about problems; the system allows uploading images as an optional feature. SeeClickFix [5] was launched in the US in 2008. These systems have not generally been adopted in Latin America, among other reasons, because they require an authority committed to take ownership for the system and respond timely to the reports. A recent analysis of six years of

FMS reports concluded that only 11% of them contain images, but also that image uploading is a significant indicator of the actual response of the authorities to the reports and the commitment of reporters to keep contributing [57]. These findings support our choice of mediating participation via geo-localized photo taking. In contrast to these systems, which by design promote individual participation [10], our work is community-oriented and puts community interests at the center.

In this sense, our work is closer to a number of open mapping initiatives in developing regions. Notable examples include the Kibera settlement in Kenya [4], and various systems built around Ushahidi [7]. In Latin America, other examples include the work done to map informal settlements in Buenos Aires, Argentina [1], and in Rio de Janeiro, Brazil [15]. Another mapping effort is led in India by Humara Bachpan [59], an organization that conducts civic campaigns centred on "child clubs" to create maps of marginalized neighborhoods. Two differences between these initiatives and our work are (1) the engagement of communities of youth in both data collection and data appropriation exercises; and (2) the development of a methodology to produce crowdsourced assessments of the conditions of photographed urban places.

Finally, social media channels are being used to generate reports of urban-related concerns, sometimes containing photos. In Mexico, Twitter has been notably used for real-time, eyewitness reports of insecurity and drug-related crimes in towns and cities [32]. This is an attractive alternative, but it is limited to people who agree to join these services and accept corporate-driven terms of use.

2.2 Crowdsourced Urban Perception

In the field of architecture and urban planning, many of the studies about visual perceptions of built environments have used qualitative methods including interviews, visual preference surveys [39, 51] and observation of the built environment using either actual or simulated images [30]. Most of these studies have been conducted and validated in either controlled laboratory settings or in actual places (which increases ecological validity). Such landscape assessment methods have been widely used and accepted [16, 27, 28, 34, 49]. In this work, we add to a growing body of work that is proposing the use of online resources like images to assess physical environments [42, 45, 50].

With the popularity of social media and mobile phones, in conjunction with an increased use of online crowdsourcing platforms to obtain judgments from diverse populations, scholars have started to explore crowdsourcing as a medium to obtain estimates of urban perception for both indoor [22, 52] and outdoor environments [42, 45, 50]. For outdoor environments, gathering perceptions typically involve the use of Google Street View (GSV) [9, 42, 50]; while GSV is widely available in the developed world, it is not so for the developing world [45]. In [50], the authors gathered geo-localized images via GSV in four developed cities in the US and Austria (called Place Pulse 1.0 dataset) and conducted a study to measure the perception of outdoor urban scenes on *safety*, *class* and *uniqueness*. In a similar study on urban perception, judgments were collected to examine visual cues that could correlate outdoor places in London with three dimensions (*beauty*, *quietness*, and *happiness*) [42]. Our current study builds upon our previous work [45], where we carried out a crowdsourcing study to collect perceptions of six dimensions of urban perception (*dangerous*, *dirty*, *preserved*, etc.) by local inhabitants of one Mexican city. Compared to [45], we collect and study data that is ten times larger and comes from three cities, define and study a larger number of urban constructs, and perform automatic inference of these constructs using visual cues extracted from images.

2.3 Automatic Inference of Outdoor Perception

To automatically infer human perception of places, most of the recent work has focused on outdoor places. Recent works have used a variety of low-level image features including Color, GIST, HOG, LBP, SIFT and more recently, generic deep convolutional activation features. Using these features on Place Pulse 1.0 dataset, high-level attributes for outdoor scenes are inferred in two US cities [33, 38]. Using the same dataset, in [41], a CNN architecture is proposed to predict and discover mid-level visual patterns which correlate with the perceived safety of an outdoor scene. Using images from Google Street View in [18], authors proposed a discriminative clustering methodology to identify visual elements (e.g. windows, balconies, and street signs) unique to the cities of Paris and London (HOG and Color features). Building upon this work, the authors in [8] proposed a scalable visual processing framework to identify the relationships between the visual appearance of a city and some of its non-visual attributes (e.g. crime statistics, housing prices, etc.). In another study to identify city-specific attributes [63], the authors conducted an analysis of 2 million geo-tagged Panoramio images from 21 cities to discover salient visual features of outdoor scenes. In a more recent study, authors in [19] expanded the Place Pulse 1.0 dataset to create a new crowdsourced data (called Place Pulse 2.0) consisting of pair-wise comparisons for over 100K images from 56 large cities across six labels: *safe*, *lively*, *boring*, *wealthy*, *depressing*, and *beautiful*. Using the collected dataset, the authors trained a CNN model to predict judgments of pairwise image comparisons by taking an image pair as input.

This paper is an extended version of our previous published paper [54], where we presented a mobile crowdsourcing methodology to collect images, conducted online crowdsourcing experiments to gather human impressions, and performed descriptive and cross-city analyses of the collected annotations. In this study, we extend our prior work by performing automatic inference of the studied urban constructs using both low-level and generic deep learning features extracted from images. We believe that developing cities need to be studied in the context of urban perception. In this regard, most other works have studied cities in developed countries, while our study focuses on an example of developing cities with different characteristics.

3 DATA COLLECTION FRAMEWORK

In this section, we describe our data collection framework including the criteria to select cities in Mexico, the definition and selection of urban perception dimensions, and the mobile crowdsourcing methodology to collect geo-localized images.

3.1 Selection of Cities

To collect images from outdoor urban spaces, we studied three small to mid-size cities in central Mexico: Guanajuato (pop. 170,000, 2005 census), Leon (pop. 1.5 million, 2010 census) and Silao (pop. 147,000, 2005 census). Guanajuato is a historical and touristic city, and the capital of a state of the same name. Guanajuato occupies a valley, forming a complex network of narrow roads, pedestrian alleys, and stairways running uphill. The city of Leon is a business and industrial hub. As per 2010 census, Leon is the seventh most populous metropolitan area in Mexico [2]. Due to its relatively larger size, some areas in Leon are quite inaccessible either due to safety concerns or because of the presence of large walls which typically surround up-scale neighborhoods. In contrast, Silao is a local hub of agricultural and industrial activity in the region, with a wide variety of farm crops, and dairy packaging plants. The three cities reflect a common situation in Latin American urbanization, which produces complex environments with historical sites, suburban sprawl, affluent neighborhoods, and informal settlements. For the three cities, images were captured from areas that included different neighborhoods reflecting the characteristics of each city, as well as touristic and historical sites. Figure 1 shows a sample of images from each city.

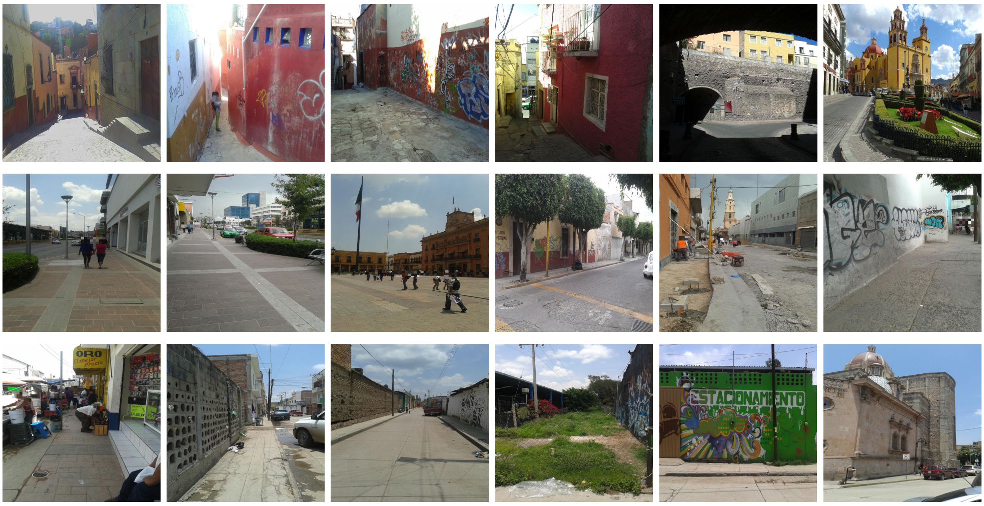


Fig. 1. Random selection of images from the *city-image* corpus. Top row shows images from the city of Guanajuato, middle row shows images from Leon, and bottom row shows a random selection of images from Silao. For privacy reasons, images showing faces have been pixelated.

3.2 Definition and Selection of Labels

In order to select labels to characterize urban perception for outdoor environments, we base our methodology on prior work [42, 45, 50]. The list of selected labels (shown in Table 1) in our study encompasses the labels studied in the literature (*accessible, dangerous, dirty, happy, interesting, picturesque, preserved, pretty, quiet, wealthy*), in addition to new ones (*polluted, pleasant*). Concretely, [50] studied *safety*, *class*, and *wealthy* dimensions; [42] examined *beautiful*, *quiet*, and *happy* labels; and in our previous work we have examined *dangerous*, *dirty*, *nice*, *conserved*, *passable*, and *interesting* [45]. We have chosen this list of labels for several reasons. First, these labels encompass physical and psychological constructs evoked while describing characteristics of the built environment. Second, all the three cities face various problems including crime, prevalence of alcohol and drugs, and streets with garbage and non-artistic graffiti, etc. These issues not only affect the well-being and safety of its citizens, but also hurt the image of a city e.g., as a tourist destination. Thus, it is essential to study the role these perceptions play in these cities. Throughout the paper, we will use the umbrella term urban perception to refer to these labels.

At this stage it is important to note that the objective of the paper is to understand urban perceptions and impressions, in the strict sense in which impressions are defined as per Brunswik's lens model, i.e., judgments made about a place based on available cues of the place (specifically, the visual cues in the observed images), which corresponds to cue utilization in the lens model [12].

3.3 Mobile Crowdsourcing Design

The images used in this study were collected using mobile crowdsourcing as part of an Urban Data Challenge (UDC). The UDC was co-designed with the aid of student volunteers (16–18 years old) from a technical school in Guanajuato city. The data challenge was carried out during a 12-week period starting in late February 2014. Student volunteers were organized into teams. Each team was given an Android-based smartphone. However, students also used their own mobile devices for data collection. To cover Leon and Silao, student teams visited these cities in person, which are 56KM and 25KM respectively from Guanajuato. We developed a mobile application that enabled students to take pictures and upload them to our image server. Mechanisms to incentivize participation included creation of study circles to raise awareness about the importance of understanding urban phenomena through the use of mobile technology, and the role of citizens in proposing creative, community-based solutions to prevalent urban problems. The UDC produced over 7,000 geo-referenced images.

In most computing research examining urban impressions, the images have been obtained using Google Street View (GSV). As argued in Section 2, while GSV provides a automated way to collect images, this approach is mostly applicable to developed cities, where GSV images are spatially comprehensive. In [45], the authors found that 58% of the images collected (via mobile crowdsourcing) were either unavailable or erroneous in the GSV database for the city of Guanajuato in Mexico due to the nature of the streets and alleys. Work using GSV imagery implicitly assumes that this is not an issue, yet it is a fundamental issue in the context of the developing world. As a result, in our study we used mobile crowdsourcing as means to collect data in a more participatory and opportunistic way.

3.4 City Image Corpus

As described above, during UDC we collected an image corpus consisting of 7,000 geo-tagged images. For our current analysis, we focus on a random selection of 1,200 images with 400 images per city, which we call the *city-image* corpus. All images were taken between 9AM and 5PM during workdays. The collected image set consists of outdoor images captured at touristic hotspots, key historical sites, traditional neighborhoods, main squares, thoroughfares, main/commercial streets and downtown areas. All the images were taken from a first-person perspective, corresponding to the natural situation in which a person navigates and perceives the urban environment. Volunteers were asked to avoid beautifying images or applying digital filters, as is usually the case with social media images, like Instagram. It is important to note that the *city-image* corpus contains not only those images that document an urban concern, but also images which capture the ebb and flow of the city while depicting different aspects of urban life and build environments. Figure 1 shows a sample of images from the corpus for each city.

Place Pulse Dataset: In addition to our *city-image* corpus, in this study we also analyze the publicly available Place Pulse 1.0 dataset [50]. As described in Section 2, the authors in [50] gathered 4,136 geo-localized images from four developed cities in the US and Austria (New York City, Boston, Linz and Salzburg). Images from NYC and Boston were gathered using Google Street View, while images in two Austrian cities were collected manually. On these images, the authors built an online platform to collect human perception ratings and measure the urban perception on *safety*, *class* and *uniqueness*. As a result, the Place Pulse 1.0 dataset contains aggregated human ratings (called Q-scores which were computed using pair-wise comparisons) for each image along the three studied dimensions. Note that the Q-score for each image and label ranges between 0 and 10. In the rest of the paper, we refer to this dataset as Place Pulse dataset.

4 CROWDSOURCING IMPRESSIONS

To gather impressions of online annotators, we designed a crowdsourcing study on Amazon Mechanical Turk (MTurk). We chose US-based “Master” annotators with at least 95% approval rate for historical HITs (Human Intelligence Tasks). In each HIT, the workers were asked to view an image of an urban space, and then rate their personal impressions based on what they saw along 12 labels. In other words, images served as stimuli to rate perceptions for 12 urban perception labels, along a seven-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (7), as typically done in psychology and urban planning research [22, 31]. Workers were required to view images in high-resolution. Workers were not given any information of the studied city to reduce potential bias and stereotyping associated to the city identity. We collected 10 annotations for each image and label, resulting in a total of 12,000 responses and 144,000 individual judgments for 1,200 images from the *city-image* corpus. Every worker was reimbursed 0.10 USD per HIT. From a methodological point of view, we differ in the way labels are collected: pairwise comparisons in prior work ([19, 38, 42, 50]) vs. individual ratings as we do in our work. We have followed an approach more commonly used in environmental psychology that allows us to compute, report, and compare standard measures of inter-rater agreement (ICC), something that is missing in previous works [19, 38, 42, 50].

In addition, we also gathered crowdworkers' demographics via an email-based survey. We asked workers about their age group, gender, level of education, current place of residence (categorized as rural, suburbs, small town, mid-size town, or city), and any experience visiting developing countries, in any region including Latin America, Asia and Africa.

4.1 Worker Participation and Demographics

For a total number of available 12,000 HITs, we observe that workers completed an average of 82 HITs, while they could potentially undertake 1,200 HITs (400 HITs per city). We had a pool of 146 workers who responded to our tasks. While 50% of the workers submitted less than 30 HITs, the worker with the highest number of HITs completed 624 assignments. We observe a long-tailed distribution in HIT completion times (mean: 59 secs, median: 43 secs, max: 297 secs). Note that we allocated a maximum of 5 minutes per HIT.

Of all 146 HIT respondents, 53% replied to our demographics survey. We notice a slightly skewed gender ratio (58% of workers being female). 80% of respondents reported their ethnicity as White/Caucasian, 12% as Asian, and 3% each belonging to Hispanic/Latino and Black/African American communities. 45% of respondents are college graduates. Furthermore, we also notice that the worker population is relatively middle age with the most popular category (43%) being the age group of 35-50 years old (18–24: 3%, 25–34: 32%, 50+: 22%). While only 18% of our worker pool reported to live in a big city, majority of them (45%) are sub-urban (for the remaining categories: rural: 18%, mid and small sized town: 9% each). Only a minority (23%) of the survey respondents reported having experience visiting any country in the developing world. For those with traveling experience in developing countries, holidays and tourism were the main purposes of the visit (55%). Amongst the visited countries, 44% of these subset of respondents have traveled to Mexico, which is not surprising given that the pool of crowdworkers are US-based. While we are not claiming that the observed trends on the crowdworker demographics are generalizable to the 146 participants who completed the HITs of our study, the observed gender ratio corroborates earlier findings in the online crowdsourcing literature [44], including a recent large-scale longitudinal study examining population dynamics and demographics of Amazon Mechanical Turk workers [17].

5 CROWDSOURCED ANNOTATIONS ANALYSIS

In this section, we first assess the quality of crowdsourced annotations, present a descriptive analysis of the aggregated annotations for each studied city, and perform correlation and PCA analysis to understand associations between the studied labels.

5.1 Annotation Quality

We begin by analyzing the annotations in terms of inter-annotator (or inter-rater) agreement. We measure the inter-rater agreement by computing intraclass correlation coefficients (ICC) among

Label	Guanajuato	Leon	Silao	Combined
Accessible	0.86	0.55	0.36	0.72
Dangerous	0.83	0.65	0.73	0.76
Dirty	0.85	0.72	0.70	0.78
Happy	0.82	0.76	0.61	0.78
Interesting	0.61	0.70	0.60	0.70
Pleasant	0.83	0.77	0.66	0.79
Picturesque	0.77	0.69	0.64	0.76
Polluted	0.68	0.56	0.57	0.64
Preserved	0.82	0.75	0.63	0.77
Pretty	0.80	0.69	0.66	0.76
Wealthy	0.84	0.73	0.57	0.76
Quiet	0.71	0.65	0.53	0.73

Table 1. $ICC(1,k)$ scores of 12 dimensions for each city. All values are statistically significant at $p < 0.01$.

ratings given by the worker pool [56]. Our annotation procedure requires every place to be judged by k annotators randomly selected from a larger population of K workers. $ICC(1,k)$, which stand for average ICC measures, are computed for each label and city across all images. Table 1 reports the $ICC(1,k)$ values for all cities for $k = 10$. In addition to listing the individual scores for each city and label, we also report the combined $ICC(1,k)$ scores for each label and the whole dataset, where we have combined all places across cities. We observe acceptable inter-rater consensus for most labels, with all values being statistically significant (p-value < 0.01).

We notice that the ICC is equal or above 0.7 for 12 of the 14 dimensions in Guanajuato, 6 in Leon, and 2 in Silao. Furthermore, only one ICC value is below 0.5 (*accessible* in Silao). This suggests that MTurk observers tend to agree on their perceptions of most dimensions. It is interesting to note that at the combined label *quiet* achieves high agreement from images not showing any sound (0.73 combined score). On the other hand, the label *polluted* is the one with lowest combined ICC (0.64). We also observe that *accessibility* has low ICC for two of the three cities. Silao has overall received the lowest ICC scores compared to the other two cities.

5.2 Descriptive Statistics

Given the multi-annotator impressions, it is necessary to create a composite score for each image, given a label. To gather the individual ratings, we used an ordinal scale, which implicitly describes a ranking. It is known that the central tendency of an ordinal variable is better expressed by the median [58]. Thus, we compute the median score for each label given the 10 responses per image. Given the median scores, we then compute the mean scores and standard deviations for each label using all 400 images for each city.

Table 2 lists the descriptive statistics for each city and label. At the level of individual annotations, the minimum and maximum values are 1 and 7 respectively for each label and city, indicating that the full scale was used by the crowd-workers. The mean scores for the majority of labels is below 4 for each city, which indicates a trend towards disagreement with the corresponding label. On the other hand, each city has urban sites that score high and low for each dimension.

In all cities, the mean scores for *accessible* are the highest amongst all labels. On all labels phrased positively (except *accessible* and *wealthy*), Guanajuato scores the highest amongst all cities, which is not surprising given that Guanajuato is a UNESCO world heritage site with a vibrant tourism industry. In contrast, *wealthy* has the lowest mean score for all cities, which is not surprising either

Label	Guanajuato	Leon	Silao
Accessible	4.62 (1.1)	5.02 (0.8)	4.41 (0.7)
Dangerous	2.92 (1.2)	2.86 (0.8)	3.17 (0.9)
Dirty	3.00 (1.2)	3.05 (0.9)	3.44 (1.0)
Happy	3.97 (1.1)	3.69 (0.8)	3.36 (0.8)
Interesting	4.38 (1.0)	3.63 (0.8)	3.50 (0.8)
Pleasant	4.13 (1.1)	3.83 (0.8)	3.48 (0.8)
Picturesque	3.55 (1.2)	3.00 (0.9)	2.73 (0.8)
Polluted	2.55 (0.9)	2.93 (0.8)	3.19 (0.9)
Preserved	4.04 (1.2)	4.00 (0.9)	3.48 (0.9)
Pretty	3.41 (1.2)	3.10 (0.9)	2.80 (0.9)
Quiet	4.08 (0.9)	3.24 (0.8)	3.10 (1.0)
Wealthy	2.58 (1.0)	2.90 (0.8)	2.43 (0.7)

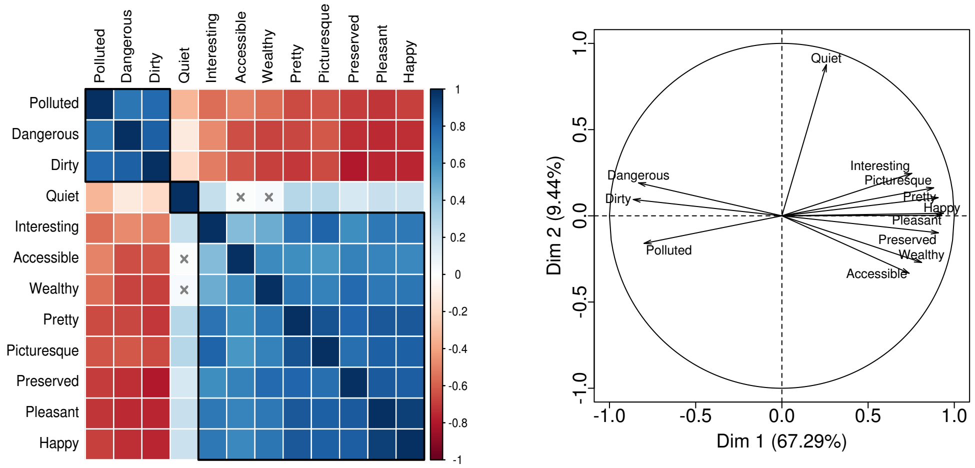
Table 2. Means and standard deviations (in brackets) of annotation scores for each city and label.

given the type of cities we are studying and the intended goals of the crowdsourced collection, leaning towards documenting urban concerns. From Table 2, we observe variation in the mean values across cities for some of the labels, but a few differences stand out. For instance, the mean differences of the *picturesque* and *interesting* attributes between Guanajuato and Silao, and the *quiet* attribute between Guanajuato and Leon and Silao all exceed 0.8, potentially suggesting differences in city perceptions. A systematic analysis to statistical testing of these differences are presented in our previous work [54].

When comparing our findings with prior work, we present results that encompass and expand all the previously studied dimensions ([50] and [38] studied 3 dimensions; [42] studied 3 dimensions; [19] studied 6 dimensions). Furthermore, the descriptive statistics in our work show urban perceptual differences with respect to work done in developed cities for the dimensions that can be compared to some degree. More concretely, the MIT Place Pulse team [50] reported Q-scores (defined to be a number between 0–10, where the middle of the scale is 5, Table 1 in [50]), while we report statistics labeled according to a Likert scale between 1 and 7 (so the middle of the scale is 4, Table 2). Besides the larger number of dimensions, one can directly compare the *wealthy* dimension (called *class* in [50]). Clearly, all developed cities have a much higher *wealthy* score compared to the developing cities. Furthermore, if we compare the *dangerous* dimension (corresponding to the opposite of *safety* in [50] and treated for the sake of argument as (10−*safety*)), we can see that the developed cities are closer to the mean of their scale than developing cities, which have lower values.

5.3 Correlation and PCA Analysis

To understand basic statistical connections between urban perception labels, we perform correlation analysis using the mean annotation scores for all labels. Figure 2a visualizes the correlation matrix across all dimensions using the aggregated data for all cities. We have used hierarchical clustering to re-order the correlation matrix in order to reveal its underlying structure. For hierarchical clustering, we adopted an agglomerative approach with complete linkage scheme. We color code the matrix instead of providing numerical scores to facilitate the discussion. We observe three distinct clusters. Starting from the bottom right in the first cluster, all the positive labels *happy*, *preserved*, *pretty*, *picturesque*, *pleasant*, *interesting* and *wealthy* are highly collinear with pairwise correlations exceeding 0.7. The second cluster consists of urban sites which are *quiet*. The third cluster (top-left) lies on the opposite spectrum with respect to cluster one, and consists of *dangerous*,



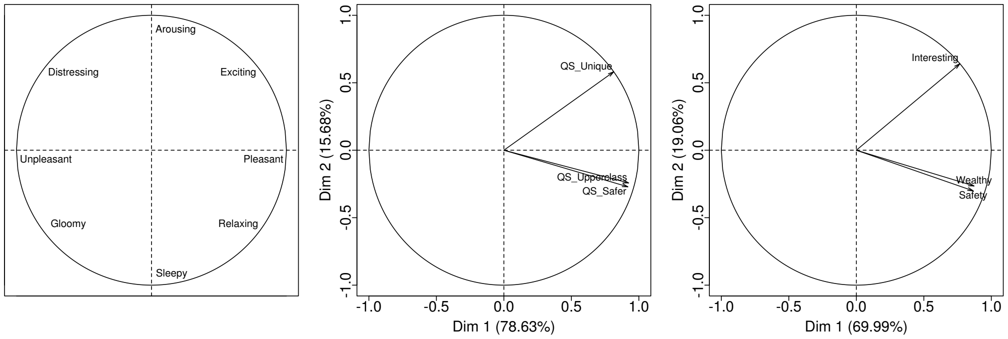
(a) (b)

Fig. 2. Correlation and PCA plots. (a) Plot showing the correlation matrix between dimensions. Matrix is color coded as per the palette shown in the right, with blue (resp. red) indicating positive (resp. negative) correlation coefficients. Black rectangular borders indicate the three distinct clusters found in the correlation matrix. Cells marked *X* are *not* statistically significant at $p < 0.01$. (b) Plot showing the first two principal components on aggregated annotation scores on 1,200 images across all cities.

dirty and *polluted*. Each of these clusters correspond to different aggregate impression, the first and third somewhat resemble "sentiment" i.e., positive/negative. As such, we can also observe significant negative correlations between dimensions in cluster one and cluster three.

To further explore the relationships between labels, we perform principal component analysis (PCA) on the aggregated annotation scores for all 1,200 images. PCA is a statistical method to linearly transform high dimensional data to a set of lower orthogonal dimensions that best explains the variance in the data [40]. In Figure 2b, we show the first two principal components which explain 77% of the variance in the annotation scores along the 12 dimensions. Note that before applying PCA, the labels were scaled to unit variance. We observe that the first component, which accounts for 67% of the variance, contains labels that resemble either the positive or negative "sentiment", respectively, on the right and left side of the X-axis. Furthermore, the second principal component primarily contains label *quiet*. These results corroborate the findings from correlation analysis and have support from early work in environmental psychology [49], as discussed below.

Revisiting the Circumplex Model of Affect for Physical Environments: In a seminal work, Russell et al. proposed a circumplex model of affect for physical environments [49], building upon their earlier research aimed at proposing a general model of affect [47, 48]. In [49], the authors proposed that the underlying structure associated with the affective experience of places can be characterized by "two orthogonal bipolar dimensions of pleasant–unpleasant and arousing–sleepy", as shown in Figure 3a. (Note that the Figure 3a has been reproduced based on the model proposed in [49].) In other words, the emotional human response to the physical environments can be mapped into a two-dimensional circular space, where the horizontal X-axis ranges from unpleasant to pleasant (i.e., valence dimension), while the Y-axis represents arousal attributes, ranging from sleepy to arousing.



(a) Circumplex Model of Affect (b) Place Pulse Dataset (c) City Image Dataset

Fig. 3. (a) Circumplex model of affect for physical environments as proposed by Russell et al. [49]. Note that this figure has been reproduced following [49], (b) Plot showing the first two principal components on the aggregated annotation scores on Place Pulse dataset, and (c) Plot showing the first two principal components using the aggregated MTurk annotations of the City Image corpus. Note that scores for the *dangerous* label have been reversed to match the *safety* label of the Place Pulse dataset.

In this paper, most of the studied dimensions are psychological in nature (e.g., pleasant, interesting, picturesque, happy, etc.). Applying PCA on the aggregated MTurk annotation scores of the *city-image* corpus and mapping the first two orthogonal components suggest a pattern similar to Russell's circumplex model, as shown in Figure 2b. We observe that the right side of X-axis contains positive valence labels i.e., pleasant, happy, pretty, and preserved, each of these labels are aligned between ~0−5°; while, the negative X-axis contains labels that evoke the affective feeling of unpleasantness (negative valence). However, it is important to note one key difference in how the stimulus was elicited between Russell et al. and our current study. In [49], participants who were physically present at venues were asked to rate the environment, while in our case images were used as stimuli to gather annotations in an online setting.

To further validate these findings, we applied PCA on the Place Pulse dataset, which was collected independent of our data collection across developed cities. Place Pulse contains impression ratings across three dimensions: *safety*, *class*, *uniqueness*. In terms of wording, these labels do not exactly match the labels studied in this paper. Figure 3b shows the first two principal components that alone explain 94% of the variance in the annotation scores along the three dimensions. In addition, we chose three labels from our study that are closer in affective meaning to the dimension studied in the Place Pulse study i.e., *interesting* (resp. *unique*), *wealthy* (resp. *upperclass*), and *dangerous* (resp. *safe*). In Figure 3c, we show the first two orthogonal components (explaining 89% of variance) of these three chosen dimensions on the *city-image* corpus. Note that the scores for the *dangerous* label were reversed to match the *safety* label of the Place Pulse study. The principal component analysis on these two datasets (Figure 3b and 3c) suggest an association between labels and provide empirical support for Russell's circumplex model of affect as applied to built environments.

6 INFERENCE

In Section 5, statistical analyses suggest that outdoor environments can be assessed in terms of reasonable inter-rater agreement for most of the urban dimensions, suggesting the presence of visual cues that allow to create such impressions. In this section, we address RQ2 to examine the feasibility to automatically infer these perceptual impressions using visual cues from images. For automatic inference we used the *city image* corpus consisting of 1,200 images across three cities.

Visual Feature	Description	Dimensionality
Color	Color histogram in RGB space	512
GIST	Dominant spatial structure of a scene	512
HOG	Histogram of oriented gradients	680
LBP	Spatial arrangement of color and intensities	256
CNN-CP	Final layer class probabilities using a GoogLeNet CNN pre-trained on *Places205*	205
CNN-FC	Fully connected layer of a GoogLeNet CNN pre-trained on *Places205*	1024

Table 3. Summary of the visual features extracted from images.

6.1 Visual Feature Extraction

To automatically infer outdoor perception, we extracted a set of low-level and deep learning visual features, building upon recent work in the literature [33, 38, 53]. Table 3 summarizes the list of visual features extracted from images.

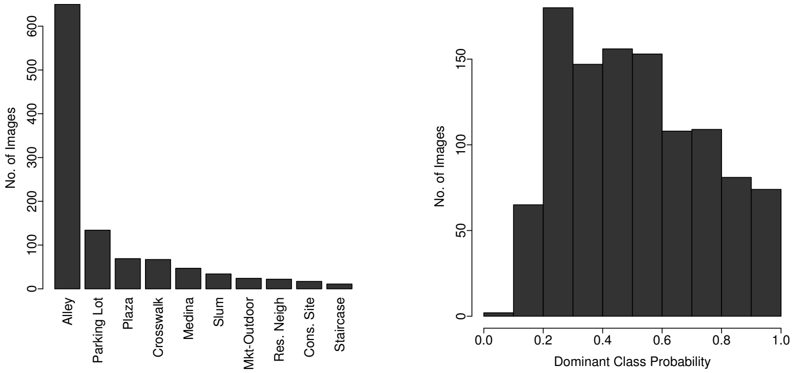
6.1.1 Low-level Visual Features. For low-level image features, we extracted color histograms in RGB space, GIST descriptors to capture the dominant structure of the outdoor scene [37], histograms of oriented gradients (HOG) [26], and LBP features which encodes local texture information [36]. Table 3 summarizes the dimensionality for each low-level feature type.

6.1.2 CNN Visual Features. : To extract deep learning features on the *city-image* corpus, we chose a pre-trained convolutional neural network (CNN) model which uses the GoogLeNet architecture trained on *Places205* database [62], in contrast to the popular *ImageNet* database. *Places205* database is a large-scale scene-centric database of 2.5 million images of indoor and outdoor scenes across 205 categories, while *ImageNet* is an object-centric database. It has been shown that *Places205* database contains more images per scene category compared to *ImageNet* [62]. Recently, authors in [41] reported that features extracted from a pre-trained Places-CNN achieve higher performance accuracy relative to a similar CNN architecture trained on *ImageNet* for outdoor scenes. Given the kind of images that was collected for the study e.g., outdoor scenes of city's built environment, residential neighborhoods, streets and alleys, etc. (see Figure 1), we believe *Places205* is more suited for outdoor places compared to *ImageNet* database.

Using the CNN model, we extracted two types of features. For the CNN-CP model, we extracted the final layer class probabilities across all *Places205* scene categories, resulting in 205-dimensional feature vector. In addition, we also extracted the output of the fully-connected layer (FC) of the GoogLeNet architecture as additional feature representation (CNN-FC model in Table 3). To extract CNN features, all images were re-sized to 256 × 256 pixels and subjected to mean image subtraction. In summary, for both CNN models, we used the same CNN architecture (GoogLeNet) trained on *Places205* database, these models differ only in terms of derived activation features.

6.2 Inference Method and Evaluation

We formulate the inference of urban outdoor perception as a regression problem where our objective is to predict aggregated human impressions using visual cues extracted from images. For regression we used Random Forest, which is a tree based supervised machine learning technique that is known to be robust towards overfitting on the training data [11]. For model validation, we performed m repetitions of a k-fold stratified cross-validation approach. For all the experiments, we set $m = 10$, and $k = 10$. After the model run and validation, we computed the mean of the evaluation metric



(a) (b)

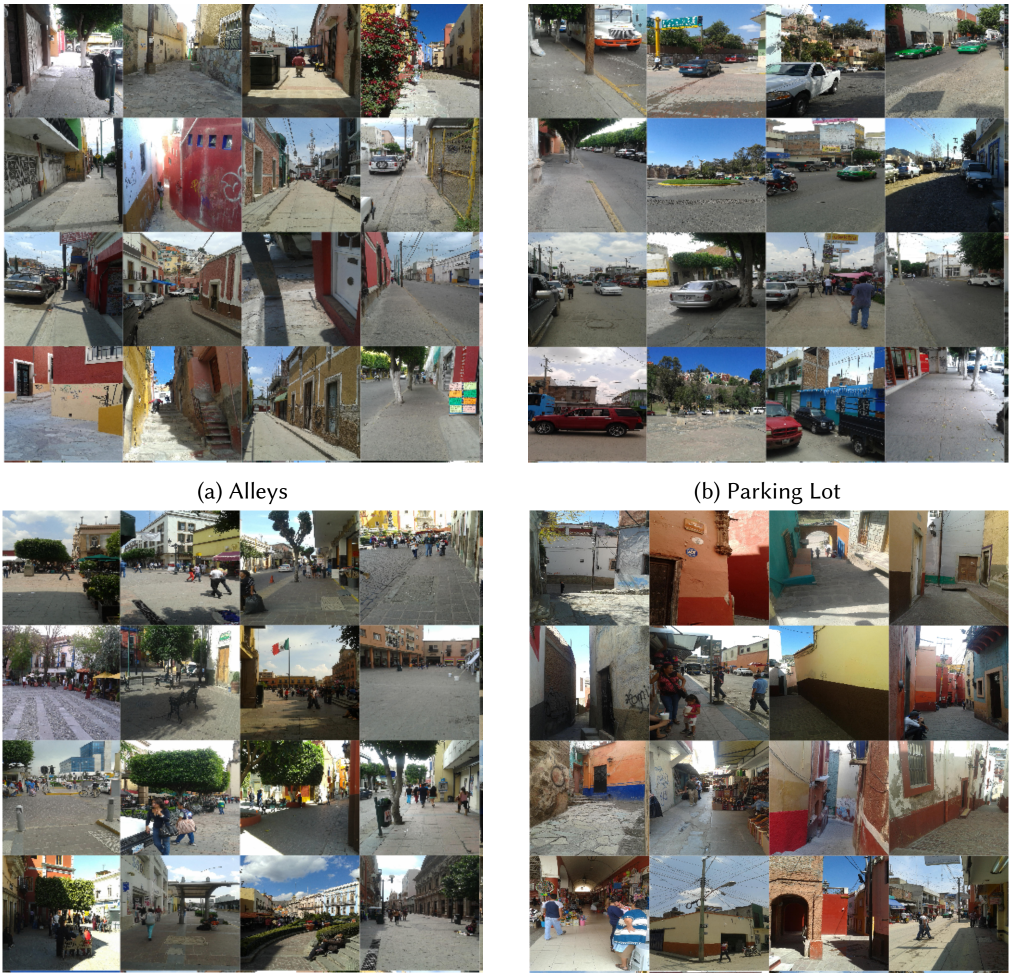
Fig. 4. a) Histogram of top-10 recognized classes for the *city image* corpus, b) Histogram of the class probabilities for the top-10 recognized categories for the *city image* corpus.

across *mk* runs as the model output. To evaluate the predictive performance between feature sets, we used root-mean-square error (*RMSE*) and coefficient of determination (R^2) as evaluation metrics. These evaluation metrics were computed between the perceived ground-truth (i.e., human judgments) and predicted perceptual scores obtained for each label and feature set. For the baseline model, we chose the mean annotated score as the predicted value for each label to better understand and compare the predictive performance of each feature set individually.

As described above, for automatic inference of urban perception, we formulated the problem as a regression problem. The other alternative was to formulate it as a classification problem. Recall that all the images were rated along a seven-point Likert scale from *strongly disagree* to *strongly agree* for each dimension (Section 4). The Likert rating scale used in our analysis has an implicit rank ordering and thus can be considered ordinal data but not necessarily interval-level data as the "distance" between successive scale items (e.g., between *strongly disagree* and *disagree* and *disagree* and *somewhat disagree*) can not be considered equivalent. Put more simply, each item category in the Likert scale can not be considered independent of each other [24]. In machine learning, many classification algorithms assume classes to be unordered, independent of each other i.e., it is assumed that there is a lack of a "natural" order amongst different class types. As a result, a typical approach to conduct predictive analysis for Likert scale or ordinal data is to consider scale items to be continuous before applying a regression algorithm. This approach has been used in the literature to quantify urban perception of both indoor [35, 53] and outdoor environments [33]. However, as part of future work, we plan to experiment with classification techniques to perform inference for ordinal datasets, that takes into account their implicit rank ordering [13, 21].

6.3 Results

6.3.1 Visual Categories. We begin our analysis by examining the distribution of the most likely *Places205* category assigned to each image. Using the CNN model, for each image in the *city image* corpus, we obtained the vector containing the class probabilities across 205 scene categories. Given this vector, we chose the scene category with the highest probability as the dominant class for each image. Figure 4a shows the histogram of the top-10 recognized *Places205* categories. To visually illustrate the recognized classes, Figure 5 shows a mosaic of images across four of the top-10 *Places205* categories from the *city image* corpus.



(a) Alleys

(b) Parking Lot

(c) Plaza

(d) Medina

Fig. 5. Sample of random images from the *city image* corpus which were classified as (a) Alley, (b) Parking Lot, (c) Plaza, and (d) Medina. For each class, from top to bottom, images are sorted in decreasing order of dominant class probability. Best viewed in color.

Overall, the dominant class distribution exhibits "long-tail" characteristics, with a total of 52 unique scene categories. 90% of the images were assigned these top-10 scene categories (out of 205 possible categories). More than 54% of the images were classified as an "alley" as their dominant class (see Figure 5a for a visual illustration). Most of the top ten dominant categories are associated with the physical attributes of the outdoor environment (e.g., alley, parking lot, plaza, crosswalk, residential neighborhood, construction site, etc.) which potentially suggests an automated way to validate the images collected as part of the *SenseCityVity* study (Figure 5). These findings further point towards the differences between images describing indoor and outdoor environments.

As a second observation, some of the recognized categories in Figure 4a may seem surprising at first glance e.g., "medina", but after manually browsing the images belonging to these categories (Figure 5d), we found that these classes describe various attributes of the outdoor environment

	Baseline	Color	GIST	HOG	LBP	CNN-FC		CNN-CP	
	RMSE	R^2	R^2	R^2	R^2	R^2	*RMSE*	R^2	*RMSE*
Accessible	0.94	0.21	0.23	0.16	0.25	**0.48**	0.69	0.46	0.70
Dangerous	0.99	0.16	0.16	0.09	0.15	**0.38**	0.79	0.35	0.80
Dirty	1.06	0.14	0.15	0.10	0.14	**0.37**	0.85	0.35	0.87
Happy	0.95	0.26	0.21	0.13	0.19	**0.46**	0.71	0.43	0.73
Interesting	0.94	0.25	0.30	0.18	0.25	**0.45**	0.70	0.41	0.73
Pleasant	0.97	0.24	0.21	0.15	0.20	**0.45**	0.73	0.42	0.75
Picturesque	1.06	0.23	0.25	0.15	0.22	**0.49**	0.76	0.44	0.79
Polluted	0.92	0.13	0.13	0.09	0.09	**0.30**	0.77	0.28	0.78
Preserved	1.05	0.20	0.20	0.15	0.18	**0.45**	0.78	0.40	0.81
Pretty	1.02	0.23	0.21	0.15	0.19	**0.46**	0.75	0.43	0.77
Wealthy	0.88	0.22	0.20	0.13	0.19	**0.46**	0.65	0.44	0.66
Quiet	1.00	0.32	0.30	0.18	0.18	**0.48**	0.73	0.45	0.75

Table 4. Inference results for 12 ambiance dimensions for all feature sets, using R^2 and *RMSE* as evaluation measures. Cells marked in **bold** correspond to the best R^2 result obtained for each dimension.

specific to the studied cities and are likely misclassified yet make sense visually. For instance, most of the images belonging to "medina" category contain images showing narrow and windings alleys and streets, which is typical to the city of Guanajuato. Of all images classified as "medina", 66% of images were from Guanajuato. The distribution of dominant class captured similarities and differences in urban characteristics across the three studied cities. For all the three cities, a similar proportion of images were labeled as *alleys*. However, more *crosswalks* and *plazas* were identified in Leon city; more *medinas* and *staircase* were recognized in the city of Guanajuato; while in Silao city, more *outdoor markets* were classified relative to other two cities. These findings corroborate the functional and urban characteristics of each city (Section 3.1).

Furthermore, we noticed that the probabilities associated with the dominant classes were not similar across images. In Figure 4b, we plot the histogram of the dominant class probabilities for the top-10 recognized scene categories. Each column in Figure 4b represents the number of images for each of the 10 probability values bins, starting with bin [0, 0.1) to [0.9, 1]. We notice that some images were classified with higher probabilities; while others were fairly difficult to classify which might be due to scarcity of *city image* corpus kind of images in *Places205* database [60]. Overall, manually browsing these images highlight the various aspects of urban life and built environment of the studied cities – narrow and winding alleys, wall graffiti, clutter of open wires hanging low in the streets, unmarked pavements, etc. Note that the UDC participants were instructed to capture images of urban scenes in their natural setting and focused on the general environment (rather than only on detected problems).

6.3.2 Predicting Urban Perception. In this subsection, we evaluate the predictive performance of both low-level image descriptors and deep learning features to infer the urban perception of outdoor scenes. Table 4 reports the R^2 values between the human impressions (i.e., ground-truth) and predicted perceptual scores for each feature set over all 12 dimensions. For the baseline model, we only report the *RMSE* values as it has $R^2 = 0$.

For automatic inference, we built two CNN models which differ in the way visual features were extracted for each image using the GoogLeNet CNN. For the CNN-FC model, image features correspond to the output of the fully-connected layer of the GoogLeNet CNN, while CNN-CP model contains the final layer class probabilities across all 205 *Places205* categories as the feature vector

(Table 3). For both the CNN models, we also report the *RMSE* values. A model having a better feature representation to estimate urban perception would result in higher R^2 and lower *RMSE* values, when compared with other models. Using the results reported in Table 4, we can make the following observations:

- CNN based features consistently outperformed low-level image descriptors (including Color Histogram, GIST, HOG, and LBP) for all urban perception labels. These findings are consistent with the results reported for indoor place ambiance [53] and computer vision literature [43].
- Overall, the results indicate that a maximum R^2 of 0.49 for *picturesque* dimension can be obtained using the CNN-FC regressor. For 9 out of 12 variables, the obtained R^2 values exceeded 0.44, with the lowest one obtained for *polluted* label.
- Among the low-level features, Color Histogram, GIST, and LBP have comparable predictive performance, while HOG performs relatively poorly for most of the dimensions. Color and GIST features achieved the best performance for the *quiet* label ($R^2 \geq 0.30$). The performance of HOG features, which have performed well for scene recognition [61], potentially suggest that the shape context within an image holds low predictive power to characterize the perceptual attributes of the outdoor environments in the *city-image* corpus.
- While examining individual labels, *polluted* achieved the lowest predictive performance across all feature sets ($R^2 \leq 0.30$), which might be associated with the labeling noise during the annotation process i.e., low inter-rater agreement (Table 1). Positively phrased labels (*happy*, *pretty*, *picturesque*, etc. – cluster 1 in Figure 2a) have higher R^2 values than negatively phrased labels (*dangerous*, *dirty* and *polluted* – cluster 3 in Figure 2a), though all the dimensions in both clusters achieve similar R^2 values.
- The *quietness* of outdoor places achieved $R^2 \geq 0.18$ for all feature descriptors, with a maximum R^2 of 0.48 using CNN-FC model. It is an interesting result as the *quietness* of a place was inferred using images alone which did not contain any form of audio sensory information. These results further corroborate findings reported in the literature for indoor places [42, 53].
- While comparing the performance of CNN-FC and CNN-CP model, we notice that CNN-FC model outperformed CNN-CP model, though the R^2 and *RMSE* values are relatively comparable for all dimensions. To visually assess the performance of the CNN-FC model, Figure 6 shows example images with the highest and lowest predicted scores for three studied dimensions.

Overall, our findings suggest the suitability of using pre-trained CNN models to infer high-level human perceptual judgments for outdoor scenes. Our findings are comparable to what has been reported in the literature. On the Place Pulse dataset, Naik et al. built a predictor combining low-level features from geometric texton histograms, GIST, and geometric color histograms to achieve an R^2 of 0.54 for the *safety* dimension [33]. Using the same dataset, the authors in [38] reported correlation coefficients (r) ranging from 0.4 to 0.7 to predict *safety*, *uniqueness*, and *wealth* using generic deep convolutional activation features. On the Place Pulse data, Porzi et al. using a different CNN architecture, reported an accuracy of 70% when predicting users' votes on image pairs for the *safety* label [41].

To further diagnose the performance of our top performing feature representation, CNN-FC model, we plot the histogram of the ground-truth and predicted values for three labels in Figure 7. Using these plots, for each dimension, we observe that the model is over-estimating lower values and under-estimating higher values i.e., even though we obtained promising R^2 values, the model is biased towards the mean value for each dimension. In other words, for images which were perceived by humans as high on *dangerous* or *picturesque*, was predicted by the model as less *dangerous* or *picturesque* and vice-versa. Recent work has proposed a CNN model that weights extreme values

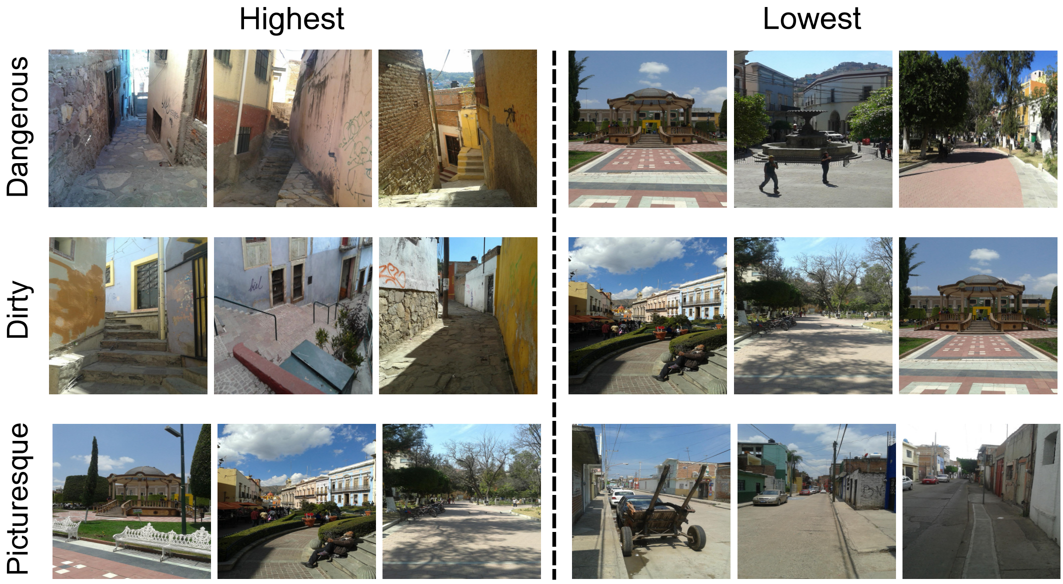
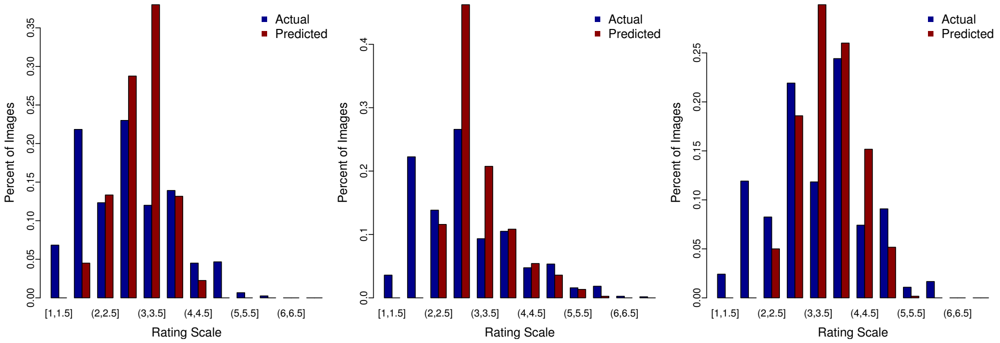


Fig. 6. Sample of images with the highest and lowest predicted scores for *dangerous*, *dirty*, and *picturesque* dimensions using CNN-FC model. Best viewed in color.



(a) Dangerous

(b) Picturesque

(c) Quiet

Fig. 7. Plots comparing the histograms of both actual and predicted values for a) Dangerous, b) Picturesque, c) Quiet.

during training as means to counter the inherent data bias due to unbalanced distributions [25]. It seems like a promising approach to increase the model performance for future work.

6.3.3 Urban Perception across Cities. Until now, we have reported and discussed the inference results using the combined data for all cities. Now, we examine the predictive performance of CNN-FC model on each city. Table 5 reports the inference results for each city across all 12 ambiance dimensions. For each city, we report the baseline *RMSE* (E_{BL}) and CNN-FC model *RMSE* (E_{RF}), in addition to reporting their respective R^2 values. Using the results listed in Table 5, we observe similar trends to the findings reported for ICC scores for each city (Table 1). Amongst all the three cities, Guanajuato achieves the best performance for all dimensions (except *interesting* and *quiet*), while Silao has received the lowest R^2 scores for most dimensions. As a second observation, the

	Guanajuato			Leon			Silao		
	E_{BL}	E_{RF}	R^2	E_{BL}	E_{RF}	R^2	E_{BL}	E_{RF}	R^2
Accessible	1.13	0.7	0.62	0.8	0.64	0.38	0.75	0.67	0.22
Dangerous	1.17	0.85	0.49	0.84	0.69	0.34	0.92	0.75	0.36
Dirty	1.22	0.91	0.46	0.93	0.79	0.29	0.96	0.82	0.31
Happy	1.11	0.77	0.54	0.81	0.62	0.41	0.8	0.67	0.33
Interesting	0.95	0.77	0.36	0.82	0.62	0.44	0.80	0.64	0.37
Picturesque	1.23	0.87	0.51	0.88	0.66	0.45	0.83	0.68	0.34
Pleasant	1.12	0.78	0.54	0.84	0.65	0.41	0.81	0.68	0.32
Polluted	0.91	0.77	0.30	0.82	0.71	0.25	0.92	0.81	0.25
Preserved	1.19	0.82	0.54	0.93	0.70	0.45	0.90	0.76	0.31
Pretty	1.18	0.84	0.51	0.87	0.67	0.43	0.88	0.73	0.33
Quiet	0.94	0.77	0.35	0.82	0.68	0.33	0.95	0.74	0.42
Wealthy	1.01	0.69	0.55	0.83	0.65	0.41	0.70	0.58	0.31

Table 5. Inference results for each city across 12 ambiance dimensions using CNN-FC model. For each city, E_{BL} and E_{RF} refer to the *RMSE* for baseline and CNN-FC model respectively.

obtained R^2 values for Guanajuato are higher when compared to the values for all cities combined for most of the dimensions (Table 4).

7 DISCUSSION

In this section, we first discuss the results of the paper, highlight the limitations of our work and suggest areas for future work. Next, we highlight the potential use of our methodology for citizens and communities in cities of the Global South.

In this paper, we presented a study to examine urban perceptions by people and machines in three cities of central Mexico as case studies of developing cities. Our study involved data collected by locals via mobile crowdsourcing, while assessments on collected data were undertaken by an external non-local population via online crowdsourcing. In other words, we have examined the perception of places as seen by "others" rather than "locals". In our study, external observers are all US-based crowd-workers. To elicit impressions of urban perceptions, the observer population plays an important role, whether the population is external (as is the case in our study) or local (local community who is familiar with the environment). It can be argued that the collected assessments by external observers induce bias in the ratings and thus limit the generalizability of our findings [55]. In our survey, most of the external observers reported not to have traveled to any developing country in the past (77% of our worker pool as reported in Section 4.1). We acknowledge this is one of the limitations of our work. However, we believe that the possibility of obtaining external perceptions of a city is valuable in and of itself to quantitatively characterize the urban landscape, especially when the cities have a large influx of visitors (i.e. travelers, tourists, students, business people, etc.) As part of future work, we plan to engage local communities to gather responses and compare their impressions with the ones obtained via online crowd annotators.

We believe that the lack of ground-truth on perceptual ratings makes it difficult to contextualize some of our findings. For most of the psychological dimensions (e.g., *happy*, *pleasant*, etc.) there exist no unique ground-truth, while for the physical dimensions (e.g. *dangerous*, *polluted*, etc.), there might be proxy measures. Previous studies have examined the relationship between the perceptions of *safety* and *class* and homicides rate in New York City [50]. However, due to the lack of publicly available data in the studied cities, such analysis was not feasible. This lack of information is a common case in developing cities. Future work could include partnerships with the city or police to investigate whether this information could be available for research. Furthermore, to contextualize some of the findings or evaluate the applications of the current work, an interesting analysis would be to gather impressions by domain "experts" (e.g., designers, architects, city planners), who are responsible for designing these urban spaces. This would facilitate the creation of a "gold standard" for visual perception research in urban places, in addition to comparing and quantifying the predictive validity of some of the proposed techniques presented in this paper.

Due to the nature of our data collection, the spatial coverage of our approach can be seen as a potential limitation. Spatial coverage includes two aspects. The first one is the spatial sampling of regions to select urban areas. We did not perform any uniform sampling to select places for our study, which we plan to do as future work for comparison purposes. The second aspect is the spatial scalability, which involves reaching diverse geographical regions. Our data collection methodology was limited to areas that could be reached by our local community. However, in the context of development, it is relevant for people to explore the urban area where they live and work in order to achieve solutions to the problems they face on a daily basis. We plan to engage other communities in the future to collect diverse datasets in other cities.

Another important aspect of our work was the use of existing pre-trained convolutional neural network models as feature extractors for our regression tasks. These deep network models, as far as we are aware, did not specifically take into account data from regions in the developing world. Yet, as we show in the paper, they are a good starting point and provide reasonable generic features that can be used for the target inference tasks. On the other hand, there are also limitations in the categorization system of these pre-trained models (e.g. see images labeled as 'Medina' in Figure 5d), which clearly point to a mismatch in content across domains. Future work would have to examine how these models could be re-trained with data that better match the typical imagery that can be seen in Latin American cities and towns, which often differs from that in the developed world. To advance research in this direction and address the lack of data from developing cities, we are publicly releasing images and annotations. We believe these resources would help to make progress on machine perception tasks that are relevant for the reality of developing cities.

Potential Impact: The research presented in this paper is part of a a larger research initiative, called *SenseCityVity*, which aims at addressing specific urban issues by young volunteers through the use mobile crowdsensing in cities of Global South, with an initial emphasis in Mexico [46]. SenseCityVity followed a interdisciplinary approach to explore the urban environment involving computer scientists and other experts on one side, and social actors (student volunteers) on the other. SenseCityVity's broader objective is to educate citizens to develop a more perceptive attitude towards realities of their urban environment and take collective action to address some of the urban civic issues in their respective local communities. We believe our work contributes to this matter (beyond scientific inquiry) by contributing tools that communities could use to generate benefits for themselves. The mobile crowdsourcing approach used to collect the data enabled participating volunteers to become more aware of their urban environment. The data collected by people provides an alternative and more comprehensive picture of the issues that matter to citizens, beyond a mapping exercise conducted by professional surveyors, which is often expensive and less detailed. To inform urban planners and design interventions in the local communities, we are currently teaming up with several NGOs and the local government to address some of the highlighted issues. Since the project inception, our methodology has been adopted by researchers in different cities of Mexico. Furthermore, in addition to geo-localized images, *SenseCityVity* project also resulted in a collection of videos of urban scenes and video-recorded interviews of locals, which was subsequently used for community reflection and artistic creation as described in [46].

8 ACKNOWLEDGEMENTS

This work has been supported by the SNSF Youth@Night and Dusk2Dawn projects and EPFL–CODEV SenseCityVity project. We thank all the members of the SenseCityVity collective. We also thank all the student volunteers for their participation and the school authorities of CECYTE for the support to our project.

REFERENCES

[1] 2017. Caminos de la Villa. https://www.caminosdelavilla.org/. (2017).

[2] 2017. Delimitacion de las zonas metropolitanas de Mexico 2010. http://www.conapo.gob.mx/es/CONAPO/Zonas_metropolitanas_2010. (2017). [Online; Accessed: 15-March-2018].

[3] 2017. FixMyStreet. https://www.fixmystreet.com. (2017).

[4] 2017. Map Kibera. http://mapkibera.org. (2017).

[5] 2017. SeeClickFix. http://seeclickfix.com. (2017).

[6] 2017. UNICEF Statistics. https://www.unicef.org/infobycountry/mexico_statistics.html. (2017).

[7] 2017. Ushahidi. http://www.ushahidi.com. (2017).

[8] Sean M Arietta, Alexei Efros, Ravi Ramamoorthi, Maneesh Agrawala, et al. 2014. City forensics: Using visual elements to predict non-visual city attributes. *Visualization and Computer Graphics, IEEE Transactions on* 20, 12 (2014), 2624–2633.

[9] Michael DM Bader, Stephen J Mooney, Yeon Jin Lee, Daniel Sheehan, Kathryn M Neckerman, Andrew G Rundle, and Julien O Teitler. 2015. Development and deployment of the Computer Assisted Neighborhood Visual Assessment System (CANVAS) to measure health-related neighborhood conditions. *Health & Place* 31 (2015), 163–172.

[10] Burcu Baykurt. 2012. Redefining Citizenship and Civic Engagement: political values embodied in FixMyStreet.com. *Selected Papers of Internet Research* 1 (2012).

[11] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[12] Egon Brunswik. 1956. *Perception and the representative design of psychological experiments.* Univ of California Press.

[13] Jaime S Cardoso and Joaquim F Costa. 2007. Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research* 8, Jul (2007), 1393–1429.

[14] Manuel Castells et al. 2014. *Reconceptualizing Development in the Global Information Age.* Oxford University Press.

[15] Will Connors. 2015. Google, Microsoft Expose Brazil's Favelas. http://on.wsj.com/1V3X2qI. (2015).

[16] Terry C Daniel. 2001. Whither scenic beauty? Visual landscape quality assessment in the 21st century. *Landscape and urban planning* 54, 1-4 (2001), 267–281.

[17] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18).* ACM, 135–143.

[18] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. 2012. What makes paris look like paris? *ACM Transactions on Graphics* 31, 4 (2012).

[19] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and Cesar A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision.* Springer, 196–212.

[20] Gustavo Esteva and Madhu Suri Prakash. 2014. *Grassroots postmodernism: Remaking the soil of cultures.* Zed Books Ltd.

[21] Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *European Conference on Machine Learning.* Springer, 145–156.

[22] Lindsay T Graham and Samuel D Gosling. 2011. Can the ambiance of a place be determined by the user profiles of the people who visit it. In *International AAAI Conference on Web and Social Media.*

[23] Kotaro Hara et al. 2013. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proc. CHI.* ACM, 631–640.

[24] Susan Jamieson et al. 2004. Likert scales: how to (ab) use them. *Medical education* 38, 12 (2004), 1217–1218.

[25] Bin Jin, Maria V Ortiz Segovia, and Sabine Susstrunk. 2016. Image aesthetic predictors based on weighted cnns. In *Image Processing (ICIP), 2016 IEEE International Conference on.* IEEE, 2291–2295.

[26] Frederic Jurie and Bill Triggs. 2005. Creating efficient codebooks for visual recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on,* Vol. 1. IEEE, 604–610.

[27] Rachel Kaplan, Stephen Kaplan, and Terry Brown. 1989. Environmental preference: a comparison of four domains of predictors. *Environment and behavior* 21, 5 (1989), 509–530.

[28] Stephen Kaplan. 1988. Perception and landscape: conceptions and misconceptions. *Environmental aesthetics: Theory, research, and application* (1988), 45–55.

[29] Stephen F King and Paul Brown. 2007. Fix my street or else: using the internet to voice local public service concerns. In *Proc. of the 1st international conference on theory and practice of electronic governance.* ACM, 72–80.

[30] Pall J Lindal et al. 2013. Architectural variation, building height, and the restorative quality of urban residential streetscapes. *Journal of Environmental Psychology* 33 (2013), 26–36.

[31] Laura J Loewen et al. 1993. Perceived safety from crime in the urban environment. *Journal of environmental psychology* 13, 4 (1993), 323–331.

[32] Andrés Monroy-Hernández et al. 2013. The new war correspondents: The rise of civic media curation in urban warfare. In *Proc. CSCW*. ACM, 1443–1452.

[33] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo. 2014. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 779–785.

[34] Joan Iverson Nassauer. 1983. Framing the landscape in photographic simulation. *Journal of environmental management* 17, 1 (1983), 1–16.

[35] L. S. Nguyen, S. Ruiz-Correa, M. Schmid Mast, and D. Gatica-Perez. 2017. Check Out This Place: Inferring Ambiance from Airbnb Photos. *IEEE Transactions on Multimedia* PP, 99 (2017), 1–1.

[36] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence* 24, 7 (2002), 971–987.

[37] Aude Oliva and Antonio Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.

[38] Vicente Ordonez and Tamara L Berg. 2014. Learning high-level judgments of urban perception. In *European Conference on Computer Vision*. Springer, 494–510.

[39] Kate Painter. 1996. The influence of street lighting improvements on crime, fear and pedestrian street use, after dark. *Landscape and urban planning* 35, 2 (1996), 193–201.

[40] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572.

[41] Lorenzo Porzi, Samuel Rota Bulò, Bruno Lepri, and Elisa Ricci. 2015. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 139–148.

[42] D. Quercia et al. 2014. Aesthetic capital: what makes london look beautiful, quiet, and happy?. In *Proc. CSCW*. ACM, 945–955.

[43] Ali Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proc. CVPR*. 806–813.

[44] Joel Ross et al. 2010. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2863–2872.

[45] Salvador Ruiz-Correa, Darshan Santani, and Daniel Gatica-Perez. 2014. The Young and the City: Crowdsourcing Urban Awareness in a Developing Country. In *Proceedings of the First International Conference on IoT in Urban Space (Urb-IoT)*. 74–79.

[46] Salvador Ruiz-Correa, Darshan Santani, Beatriz Ramirez-Salazar, Itzia Ruiz-Correa, Fatima Alba Rendon-Huerta, Carlo Olmos-Carrillo, Brisa Carmina Sandoval-Mexicano, Angel Humberto Arcos-Garcia, Rogelio Hasimoto-Beltran, and Daniel Gatica-Perez. 2017. Sensecityvity: Mobile crowdsourcing, urban awareness, and collective action in mexico. *IEEE Pervasive Computing* 16, 2 (2017), 44–53.

[47] James A Russell. 1978. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology* 36, 10 (1978), 1152.

[48] James A Russell. 1979. Affective space is bipolar. *Journal of personality and social psychology* 37, 3 (1979), 345.

[49] James A Russell et al. 1980. A description of the affective quality attributed to environments. *Journal of personality and social psychology* 38, 2 (1980), 311.

[50] Philip Salesses, Katja Schechtner, and Cesar A Hidalgo. 2013. The Collaborative Image of The City: Mapping the Inequality of Urban Perception. *PLoS ONE* 8, 7 (07 2013), e68400.

[51] Robert J Sampson et al. 2004. Seeing disorder: Neighborhood stigma and the social construction of "broken windows". *Social psychology quarterly* 67, 4 (2004), 319–342.

[52] Darshan Santani and Daniel Gatica-Perez. 2015. Loud and trendy: Crowdsourcing impressions of social ambiance in popular indoor urban places. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 211–220.

[53] Darshan Santani, Rui Hu, and Daniel Gatica-Perez. 2016. InnerView: Learning Place Ambiance from Social Media Images. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 451–455.

[54] Darshan Santani, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2015. Looking at cities in Mexico with crowds. In *Proceedings of the 2015 Annual Symposium on Computing for Development*. ACM, 127–135.

[55] Darshan Santani, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2017. Insiders and Outsiders: Comparing Urban Impressions between Population Groups. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 65–71.

[56] Patrick E Shrout et al. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420.

[57] Fredrik Matias Sjoberg et al. 2015. The Effect of Government Responsiveness on Future Political Participation. *Available at SSRN 2570898* (2015).

[58] S. S. Stevens. 1946. On the Theory of Scales of Measurement. *Science* 103, 2684 (1946), 677–680.

[59] Sam Sturgis. 2015. Kids in India Are Sparking Urban Planning Changes by Mapping Slums. http://bit.ly/1LtS9S4. (2015).

[60] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1521–1528.

[61] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 3485–3492.

[62] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.

[63] Bolei Zhou, Liu Liu, Aude Oliva, and Antonio Torralba. 2014. Recognizing city identity via attribute analysis of geo-tagged images. In *European conference on computer vision*. Springer, 519–534.