



**LOUD AND TRENDY: CROWDSOURCING
IMPRESSIONS OF SOCIAL AMBIANCE IN
POPULAR INDOOR URBAN PLACES**

Darshan Santani Daniel Gatica-Perez

Idiap-RR-01-2015

JANUARY 2015

Loud and Trendy: Crowdsourcing Impressions of Social Ambiance in Popular Indoor Urban Places

Darshan Santani, and Daniel Gatica-Perez

Abstract—There is an increasing interest in social media and ubiquitous computing to characterize places in urban spaces beyond their function and towards psychological constructs like *ambiance*, i.e., the impressions people form about places when they first visit them - energetic, bohemian, loud, artsy, and so on. In this paper, we study whether reliable impressions of the *ambiance* of indoor popular places can be obtained from images shared on social media sites like Foursquare. Our contributions are two fold. First, using more than 50,000 images collected from 300 popular indoor places across six cities worldwide, we design a crowdsourcing experiment on Mechanical Turk to assess the suitability of social images as data source to convey place *ambiance* and to understand what type of images are most appropriate to describe *ambiance*. We demonstrate that images with clear views of the environment are more informative of *ambiance* than other image categories. Second, based on these results we build an image corpus of 900 images and used them to design a second crowdsourcing study to assess how people perceive places socially from the perspective of *ambiance* along 13 dimensions. We show that reliable estimates of *ambiance* can be obtained using user-contributed Foursquare images for several of the investigated dimensions, suggesting the presence of strong visual cues to form place impressions. Furthermore, we investigate whether there are any statistically significant difference across cities along *ambiance* dimensions, and found that most aggregate impressions of *ambiance* are similar across popular places in all studied cities. To the best of our knowledge, our work presents the first results on how images collected from social media sites relate to the crowdsourced characterization of indoor *ambiance* impressions in popular urban places.

Index Terms—Ambiance, Crowdsourcing, Foursquare

I. INTRODUCTION

CITIES are unique expressions of human activity and, at their core, are the intersection of physical spaces and the people who live in them. Cities are not only buildings and roads, but also the people who use them and create new knowledge and innovate through the continuous exchange of ideas and intermingling [1]. From a city perspective, public places have played a central role in facilitating a socio-cultural habitat for people to counterbalance the grind of daily life, an environment away from home and work [2], [3]. There is growing interest in urban studies on one hand and psychology on the other one to contextualize the understanding of urban spaces according to the perceptions of their inhabitants, which are rooted in both socio-economic factors and psychological constructs [4]. In those domains, the literature has started to

examine connections between psychological features of cities and key indicators like well-being and prosperity [5].

In this context, an area of active research is the development of “a better idea of how people perceive and experience places” [4]. As soon as we step into an indoor place, we can tell if it is made for us (or not). We ubiquitously judge restaurants, cafes, or bars according to their social *ambiance* – whether the atmosphere is energetic, bohemian, loud, trendy and so on. In other words, we form place impressions [6] combining perceptual cues that involve most senses as well as prior knowledge of both the physical space and its people. We use these impressions to make decisions that have long-term impact, defining our favorite hangouts and shaping our new discoveries. In the rest of the paper, when we say “places”, we refer to indoor public places such as restaurants, bars, clubs, coffee joints, etc.

One of the key challenges to understand how people perceive a place is the difficulty in obtaining place impressions. An obvious way is to physically visit a place and gather impressions making silent observations about its atmosphere [7]. Clearly, this approach is not scalable and does not capture the temporal aspects of venues e.g., a place that might be ideal for a business lunch, but that turns into a trendy loud bar in the night. Another approach is to gather place impressions based on images shared on social media, where observers rate *ambiance* after viewing user contributed images for that place. This approach has the advantage of being scalable and can easily span national boundaries to help examining spatial and cross-cultural differences in place *ambiance*, if they exist. In addition, this approach also permits a better understanding of a place based on images taken during different times of the day. In this work we conduct our *ambiance* analysis based on the later approach.

Due to the growth of sensor-rich mobile devices in the last five years, online yellow-page directory services like Yelp and Foursquare have risen at a fast pace. These platforms provide users the functionality to search for places in a given geographic region and to leave feedback in the form of *reviews* and *comments* about their personal experiences [8]. In addition, users also typically upload photos taken at the venue and share them publicly. As a result, hundreds of thousands of images illustrating different places across the globe are shared via social media channels. While many of the shared images are either personal or show food or drink related items, there are also images which provide views of the indoor environment and likely adequate to gauge the place atmosphere. Our work addresses two research questions:

RQ1: What types of social media images best convey the *ambiance* of popular indoor places?

D. Santani and D. Gatica-Perez are affiliated jointly to Idiap Research Institute, Martigny, Switzerland, and École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland (Email: dsantani@idiap.ch; gatica@idiap.ch)

RQ2: Can the ambiance of a popular indoor place be reliably assessed by observers of social media images? If so, for what dimensions of ambiance?

In order to obtain the images for our study, we base our current analysis on Foursquare as it exposes a API (Application Programming Interface¹) to obtain the user-contributed images for every place on Foursquare. After the image collection, we employ crowdsourcing as means to characterize place ambiance impressions. Recent research in both computer science and psychology confirm the feasibility of using crowdsourcing to conduct behavioral studies when appropriate incentives and mechanisms for quality control are established [9]. We use Amazon’s Mechanical Turk (MTurk) to design the crowdsourced tasks, in which workers view images and answer questionnaires based on our research goals.

Our paper has four research contributions:

- 1) Our first contribution is a collection of images of popular places from six metropolitan cities worldwide. We collected images from popular places in three world regions – North America (New York City, Seattle, and Mexico City), Europe (Barcelona and Paris) and Asia (Singapore). In addition to cultural and geographic diversity, these cities are chosen because of their active user population on Foursquare. Importantly, note that the focus of our study is on **popular indoor** places in Foursquare rather than on arbitrary places, which might or might not be representative in social media sites. Our image corpus contains more than 50,000 images across 300 places.
- 2) As a second contribution, we design a crowdsourcing experiment on MTurk to assess the suitability and quality of images which are most appropriate to describe the ambiance of a place. Using statistical tests, the results clearly show that images with clear views of the environment are more informative of ambiance than other image categories, like food and drinks, or groups of people. Based on the crowdsourcing results, we build a final corpus of 900 images (3 images per place) which are suitable for indoor ambiance characterization.
- 3) A priori, the social ambiance of places is not known to zero-acquaintance visitors or observers. As a third contribution, we design a second crowdsourcing experiment to assess how people perceive places socially from the perspective of ambiance. We asked crowdworkers to rate indoor ambiance along 13 different physical and psychological dimensions where images served as stimuli to form place impressions. The ambiance categories include romantic, bohemian, formal, old-fashioned, trendy, etc.
- 4) Based on the results obtained from the second crowdsourcing experiment, we show that reliable estimates of ambiance can be obtained using user-contributed images, suggesting the presence of strong visual cues to form accurate place impressions. Interestingly, while we identified a few statistically significant differences across cities along four ambiance dimensions, most aggregate impressions of ambiance are similar across popular places in all cities.

Our work contributes to the generation of resources that

in the future allow to investigate multimedia approaches to recognize social perception of urban venues.

The rest of the paper is organized as follows. We begin with a review of related work. Next, we describe our methodology to select the list of popular places and identify the most adequate image categories for ambiance characterization. We then investigate whether reliable estimates of indoor ambiance can be obtained using images along 13 different physical and psychological dimensions. Finally, we conclude with a summary of our findings and possible research directions for future work.

II. RELATED WORK

Given the multifaceted nature of our research questions, we review the related work along four axes: ubiquitous and multimedia computing, social media, hospitality research, social psychology and urban computing.

The existing work on place characterization in ubiquitous computing, computer vision, and audio processing, has examined several aspects including physical properties of places like their geographic location [10]; place composition, including the scene layout and the objects present in the scene [11]; place function, i.e., home, work, or leisure places; and place occupancy and noise levels [12]. This research has used both automatic [13], [14], [15] and semi-automatic approaches [16] and a variety of data sources often studied in isolation, including images, sensor data like GPS/Wifi, and RF data often using smartphone platforms. Works like [13], [15] have used audio or audio-visual data to characterized places through phone apps. The studied place categories (personal places in [13], home, work et al. in [15]) differ significantly from ours. A very recent work [12] investigates the recognition of physical ambiance categories (occupancy, human chatter, noise and music levels) using standard audio features collected in-situ by users. In contrast, our work examines social images as source of data, impressions of people who are not physically at the places, and a much larger number of social ambiance categories.

The emergence of social multimedia platforms, that allow users to take and share photos using mobile devices, have gained wide spread adoption. In the social multimedia literature, the work in [17] studies the problem of recommending locations based on mobility traces extracted from GPS and social links, without using image information. In contrast, the work in [18] uses geo-localized images, travel blog text data, and/or manual user profiles to suggest trips. Other works involving social images include [19] and [20]. These cases are focused on outdoor places and do not address the atmosphere dimensions we studied here. Due to the availability of large amount of images on Flickr or Instagram, researchers have analyzed these platforms (recent examples include [21], [22]). In [22], using a corpus of 1 million Instagram images, the authors studied the relationship between photos containing a face and its social engagement factors and found that photos with faces are more likely to receive *likes* and *comments*. As it relates to our work, an interesting result is that only 20% of images were found to contain faces, which suggests

¹<https://developer.foursquare.com/>

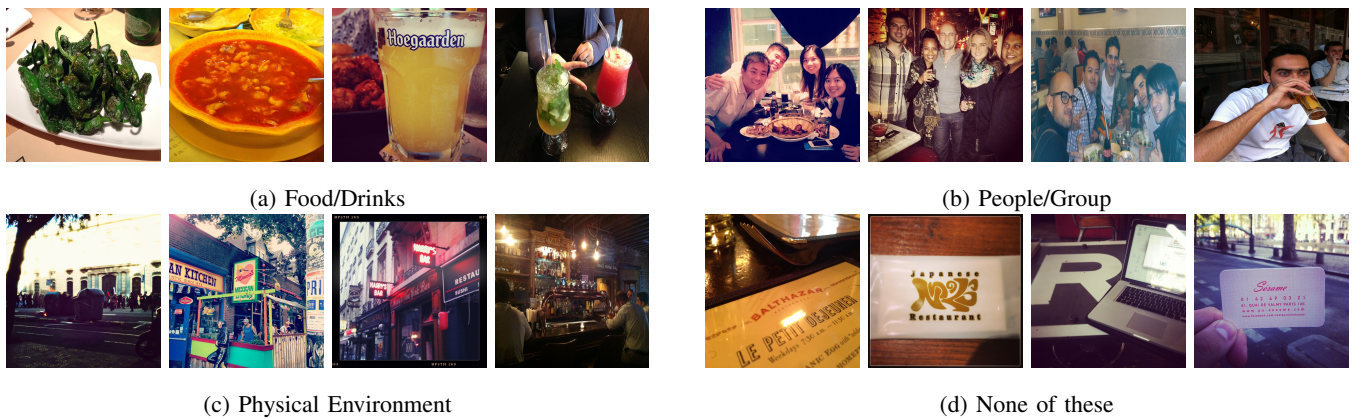


Fig. 1: Sample images from *Random-Image* corpus. Based on MTurk annotations, a random set of four images which were classified as (a) Food/Drinks, (b) People/Group, (c) Physical Environment, and (d) None of these. For privacy reasons, images showing faces will be blurred in the final version, but are kept not blurred for the reviewing process.

that many other categories of images (related to food, places, etc.) exist (for an example, refer to a small-scale coding study in [21]). To the best of our knowledge no research has been conducted so far using images shared on Foursquare for ambiance characterization, as we do here.

In hospitality and retail studies, there has been significant research interest to examine the effect of physical ambiance cues or “atmospherics” such as color, lighting, layout, style and furnishing on the customer perception and quality inferences [23], [24], [25]. In a study conducted in a retail store, Baker et al. found that ambient (such as music, lighting, smell), design (such as color, ceilings, spatial layout) and social factors present in the store environment contribute towards higher merchandise and service quality [23]. In another study, the authors in [26] investigated the role of atmospherics across 10 full-service restaurants in Hong Kong. Using five dimensions of restaurant atmospherics (facility aesthetics, ambience, spatial layout, employee factors, and the view from the window) the authors found these dimensions to have a significant influence on patrons’ dining experience, and their willingness to pay more and recommend the restaurant to others. Similar results were obtained in another related work conducted across ethnic restaurants in the U.S. [27]. However, most of these studies are either controlled laboratory settings or based on questionnaires, which may have limitations with respect to the ecological validity or recall biases.

Unlike the above research, we take a new direction examining the social perception of places (the ambiance impressions that people form about venues). Our proposed research is most closely related to work in social psychology [5], [6] which has investigated first impressions of places in connection to the personality of their inhabitants, mostly in controlled settings. A key first study [7] investigated the reliability (in terms of inter-rater agreement) of impressions of place ambiance and patron personality formed by (a) observers physically present at a number of indoor places, and (b) observers of Foursquare *user profiles* who visited those places (as opposed to views of places as we do). The results suggest that people do indeed form consistent impressions of ambiance and patrons traits. The study, however, only examined 49 places in one city

City	Ratings	Photos	Visitors
Barcelona	8.66 (0.67)	309.58 (383.53)	1,874.34 (2,371.43)
NYC	9.31 (0.41)	463.62 (387.31)	8,272.16 (6,208.76)
Paris	8.55 (0.63)	220.98 (254.16)	1,685.76 (1,433.14)
Seattle	8.95 (0.38)	240.7 (147.94)	3,533.54 (1,815.34)
Singapore	8.29 (0.86)	304.88 (206.58)	3,457.64 (3,916.89)
Mexico City	8.78 (0.49)	361.34 (374.85)	3,692.56 (3,578.84)

TABLE I: Summary statistics of Foursquare data. For each city, mean scores of attributes of popular places is shown, along with their standard deviations (shown in brackets.)

(Austin, TX) and involved personally visiting every venue. In contrast, we study places in six cities.

A recent urban computing study [28] measured the perception of three variables (safety, class and uniqueness) using 4,136 geo-tagged *outdoor* images in four cities – two in the US (Boston, New York), and two in Austria (Linz and Salzburg). Images for the US cities were obtained via Google Street View, while manual collection was performed for the Austrian cities. The authors claimed clear differences in the range of perceptions of these dimensions between American cities and their Austrian counterpart. In a similar study on *outdoor* urban perception, judgments from over 3000 individuals were collected to study visual cues that could correlate outdoor places with three dimensions (beauty, quietness, and happiness) [29]. In both studies, dedicated websites were used to collect annotations, as opposed to common crowdsourcing platforms like MTurk. Our research differs significantly from these previous works on two specific grounds. First, we are interested in examining indoor places as opposed to outdoor spaces. These are clearly different categories from the urban design and urban studies perspectives. Second, we study 13 dimensions of social ambiance (including artsy, bohemian, loud, trendy, romantic), appropriate for the indoor setting, as opposed to [28], [29], which reflect pedestrian or street-level characteristics.

III. SELECTION OF PLACES AND IMAGES FOR AMBIANCE

In this section we describe our methodology to select the list of popular places and their associated images across six cities. We then present the first crowdsourcing experiment that aims at identifying the most adequate image types for ambiance characterization and analyze the results.

A. Place Selection (Place Database)

We ground our analysis on data collected from Foursquare, which allow users to search nearby places, “locate others and be located” [30]. In Foursquare, users typically visit a place, announce their arrival (*check-in*) and share information about their visits to places with their friendship circle. As per Foursquare rules, a place or a venue is a geographical location with fixed spatial coordinates, i.e., latitude and longitude, where people can meet in person. Throughout this paper, we will use place and venue interchangeably in the context of Foursquare.

Each place on Foursquare maintains a profile page, which contains general information about the place (address, directions, phone number, etc.), in addition to Foursquare-specific data such as its popularity (on a scale from 1 to 10), total number of check-ins and past visitors, and a collection of images uploaded by users. Foursquare allows developers to obtain most of this information using its public API, which we used to gather all the relevant information for a given place.

For our analysis, we studied popular places on Foursquare for six cities around the world – Barcelona, Mexico City, New York City, Paris, Seattle and Singapore. These cities were chosen for two reasons. First, they all are large cities in diverse world regions, and are known to have a vibrant urban life. Second, they all have an active user population on Foursquare. For each city, we chose the city’s 50 most popular places which fall under the Foursquare-defined category of either being “Food” or “Nightlife Spots”, which means cafes, restaurants, or bars. Table I lists the mean values of Foursquare data for all 50 places in each city. As stated in the Introduction, we are more focused on studying popular indoor places as opposed to arbitrary places (indoor or outdoor, that might or might not be represented on Foursquare.)

Place selection was performed manually by the first author, taking into account place popularity, number of checkins, number of past visitors and number of available images. As image quality was an important criterion while selecting a place, we ignored all popular places which did not had any good-quality images such as dark images. In Table I, using data obtained from Foursquare API, we notice that the user-generated mean rating of places selected for our study is above 8.2 for all cities, confirming the popularity of places. Moreover, we also observe that these places are frequently visited by a large visitor population.

B. Image Selection

The second important consideration was the selection of images for each chosen place. We decided to select a small number of images per place to illustrate the place’s atmosphere. This decision was motivated by the need to account

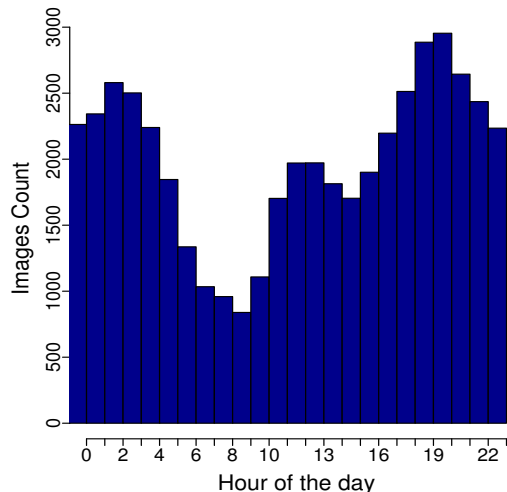


Fig. 2: Histogram showing the frequency of images taken during different times of the day.

for the variability in image quality, while at the same time providing a wide view of a place, without complicating the annotation process. Moreover, having more than one image for a place gave us the flexibility to show the place at different times of the day.

1) *Selection of Random Images (Random-Image Corpus):* Given that each place in our database has on average more than 300 user-contributed images (see Table I), it is challenging to select a small number of representative images which can accurately describe the ambiance of a place. One approach is to randomly select them given the collection of all images available for a place. We follow this approach to build a *random-image* corpus.

Images for a place listed on Foursquare can be obtained via the API, but due to rate limits imposed by the API, we have access to at most 200 publicly visible images per place. We gathered a total of 50,023 images for all 300 places. This gives an average of 166 images per place, which is lower than the average estimated from the profile metadata (≥ 300), yet it remains a large number. In addition to gathering the images, the API also provides information on the image source (i.e., the application used to generate the photo), creation time, and other attributes such as image height and width. However, due to API restrictions we had access to metadata information for only 47,980 images.

Using this information, we found that the median height and width of an image in our collected corpus is 720 pixels, with 55% of images taken via iPhone, 19% via an Android device, and 22% uploaded via Instagram. We also plot the distribution of image creation times in Figure 2. We identify three distinctive peaks – the first one occurs during the lunch hour (11am–1pm), the second peak around dinner time (6–8pm), and the last one occurs after midnight and early hours of the morning (nightlife). This result confirms our intuition that social media images provide a well-suited medium to capture places during different times of the day.

For the study described in this section, since we are interested in only a few images to represent each place, we



Fig. 3: Sample images from *Manual-Image* corpus. Based on MTurk annotations, images which scored the **highest** on (a) Artsy, (b) Creepy, (c) Loud, and (d) Trendy; and **lowest** on (e) Artsy, (f) Creepy, (g) Loud, and (h) Trendy. For privacy reasons, images showing faces will be blurred in the final version, but are kept not blurred for the reviewing process.

randomly sampled three images per place from our corpus, to build a *random-image* corpus of 900 images for all 300 places. Refer to Figure 1 for a sample of selected images.

2) Selection of Manual Images (*Manual-Image Corpus*):

Our second approach to build an image corpus is manual selection of images, where a small number of images per place are chosen manually.

The selection was performed by the first author after browsing through all the user-contributed images. During the process, we opted for images with an indoor view clearly showing the space from different angles (with or without the presence of visitors.) To the best of our ability, we avoided images where one can potentially identify faces, to protect the privacy of individuals. Moreover, we ensured that images that showed the venue name or any other information that explicitly revealed the identity of the place were ignored e.g., an image showing Starbucks or Hard Rock Cafe logos, to reduce any potential bias while characterizing the place ambiance. See Figure 3 for a sample of selected images from this corpus. Note that all these attributes cannot be controlled for while choosing images randomly.

The *manual-image* corpus also contains 900 images for all 300 places (three images per place). The manual selection was constrained by the quality of images uploaded on Foursquare. At times, we encountered images which were not optimally bright or clear. However, this setting is realistic due to the absence of any beautified or vendor-provided images, which can potentially add biases to the perceived impressions.

C. Crowdsourcing Labeling of Image Ambiance Quality

In this section we address RQ1, i.e., we use both image corpora to judge which approach results in better “ambiance quality”, that is, images which are more adequate to convey ambiance.

On one hand, random selection of images is a realistic “in the wild” setting that provides an automated way to collect images. However, it will represent a valid approach only if it results in a collection of images which provide sufficient visual cues to characterize place ambiance. On the other hand, the manual selection of images is a controlled setting that satisfies the criteria described in the previous paragraphs.

Our hypothesis is that majority of images in the *random-image* corpus is not ideal to characterize a place’s ambiance, as they do not contain visual cues to gauge a place’s physical environment. In our exploratory inspection, most of these random images contain food items or show groups of people. To answer RQ1, we designed and conducted a crowdsourcing study to gather the ability of images, both *random* and *manual*, to describe a place’s ambiance and physical environment. For crowdsourcing, we used MTurk and chose US-based workers with at least 95% approval rate for historical HITs (Human Intelligence Tasks). In addition, to increase the potential reliability of MTurk annotations, we only chose “Master” annotators, which typically involves a worker pool with an excellent track record of completing thousands of tasks with precision.

For each HIT annotation task, we picked a set of five images per place, consisting of two from the *manual-image* corpus, and three from the *random-image* corpus. We ensured that images from the two sets do not overlap for this experiment.

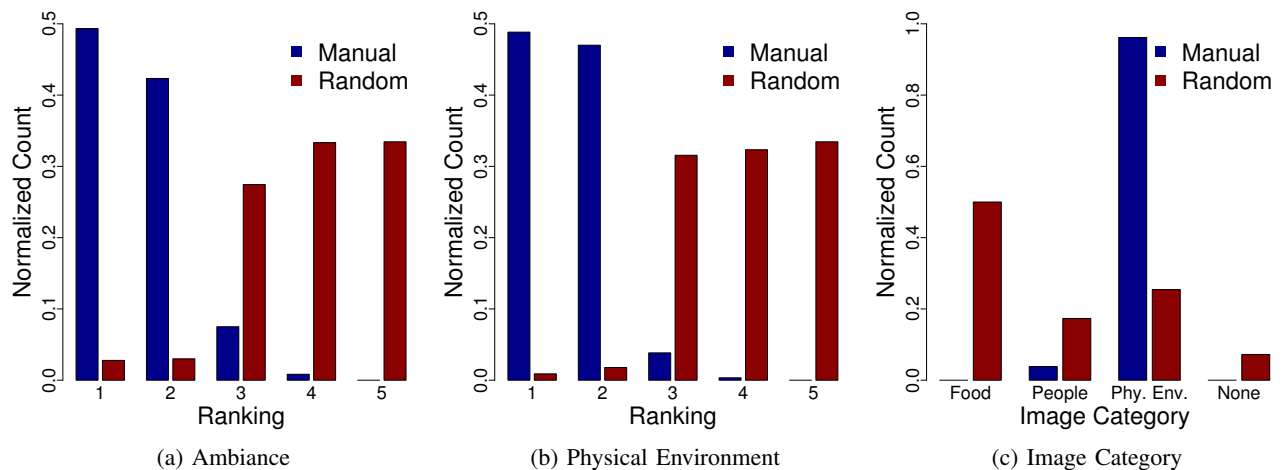


Fig. 4: Results for *Majority Vote* aggregation technique. Plot showing the histograms for a) Ambiance, b) Physical Environment, and c) Image Category, for both *Manual-Image* and *Random-Image* corpus.

In each HIT, workers were asked to view these five images and then answer three questions. In the first question, the workers were asked to rank the images based on how informative they are of the *ambiance* of the place. In the second one, workers were asked to rank the same set of images on their degree of information about the *physical environment* of the place. The third question asked workers to categorize the images in four classes: a) Food/Drinks, b) People/Group, c) Physical Environment, and d) None of these. For these questions, no explicit definitions of ambiance, physical environment, food/drinks or people were provided, as we wanted the workers to rely on their internal representation of these concepts.

In the two ranking questions, images cannot be given the same rank, each image needed to have a different rank. For the image categorization task, the workers were asked to choose exactly one category for each image. If the images had the same rank or fell into one or more categories, we asked the annotators to pick the rank or category that for them was the best choice. We collected 10 annotations for each HIT across all 300 places, for a total of 3000 responses for every question. Every worker was reimbursed 0.15 USD per HIT.

We also gathered crowdworkers’ demographics via an email-based survey. We asked workers about their age group, gender, education level, current place of residence (categorized as either rural, suburbs, small-sized town, mid-sized town or city), and any experience of living in a big city. We also inquired them about their typical frequency to go out for food or drinks (almost every day, 2-3 times per week, once a week, 1-2 times a month, or less than once a month). These questions were designed to understand the ability of workers to rate images for ambiance and physical environment based on previous experiences.

D. Results

In this section, we present the results of our first crowdsourcing experiment.

1) *Participation levels and Completion tasks*: For a total number of 3,000 HITs available for this experiment, we observe that a typical worker completed an average of 39

HITs. While 50% of the workers submitted less than 9 HITs, the worker with the highest number of HITs completed 295 assignments. We also observe a long-tailed distribution in HIT completion times (mean: 114 secs, median: 88 secs, max: 593 secs). It is worth noting that we allocated a maximum of 10 minutes per HIT.

2) *Worker Demographics*: We had a pool of 101 workers who responded to our HITs. Of all HIT respondents, 32% replied to our demographics survey. We notice a balanced gender ratio (50% of workers being female), which corroborates earlier findings in the literature [31]. While only 34% of our worker pool currently lives in a big city, 75% of them have already experienced city living in the past. Furthermore, 56% of them go out for foods or drinks at least once a week. These findings provide evidence that the majority of the responders are likely capable to assess image ambiance and form accurate ambiance impressions in urban environments. Furthermore, we also notice that the worker population is relatively not so young with the most popular category (53%) being the age group of 35-50 years old. Note that the worker demographics reported here encompasses the worker population in both crowdsourcing experiments of this paper.

3) *Image Ambiance Quality Annotations*: Now we turn our focus towards assessing the quality of image ambiance annotations. As mentioned in the last section, for each HIT we collected 10 impressions per place. Thus it becomes crucial to consider the role of different aggregation methods in analyzing the results. Aggregation is used to create a composite score per place given the 10 responses for each question. In other words, for every question, aggregation is performed at the place-level. We use two different aggregation techniques. The first one is the *majority vote*, where we compute the majority score given the 10 impressions for each place. We then summarize results based on 300 majority impressions. For the *median* method, we compute the median as the composite score across the 10 impressions for each place.

Table II lists the summary statistics for the two aggregation techniques. For each aggregation technique and each corpus, we report the percentage of images which are in Top 2 ranks

Method	Manual				Random			
	Top 2		Bottom 3		Top 2		Bottom 3	
	Ambiance	Phy. Env.	Ambiance	Phy. Env.	Ambiance	Phy. Env.	Ambiance	Phy. Env.
Majority Vote	91.7%	95.8%	8.3%	4.2%	5.8%	2.7%	94.2%	97.3%
Median	89.7%	94.7%	10.3%	5.3%	4.1%	2.0%	95.9%	98.0%

TABLE II: Table showing the summary statistics for each aggregation method. For each method, we show the percentage of images from both image corpora which are either in Top 2 ranks (rank 1 or 2), or ranked in Bottom 3 (ranks 3,4,5).

(ranks 1,2) and Bottom 3 ranks (ranks 3,4,5). We list these statistics for both the *ambiance* and *physical environment* questions. For the *majority vote* technique, manually selected images are in Top 2 ranks 91.7% for ambiance and 95.8% for physical environment, while random images are in Bottom 3 ranks for 94.2% and 97.3%, respectively. Note that a random ranking method would assign the manually selected image in the Top 2 rank with a probability of $1/10$ ($1/\binom{5}{2}$). We also plot the histogram of rankings for image sets from both corpora in Figures 4a and 4b. These results show that manually selected images are associated with higher ranks, while the random set of images are associated with lower ranks for both ambiance and physical environment, irrespective of the aggregation technique.

In addition to asking annotators to rank images, we also asked them to classify images into one of four categories, as explained in the previous section. In Figure 4c, we plot the assigned category for the *majority vote* technique. We observe that images from the *manual-image* corpus describe the physical environment in 96.2% of the cases. In contrast, images from the *random-image* corpus are representative of either food/drinks, or people, or other categories in 74.6% of cases combined, and showing food items or people in 67% of cases combined.

Statistical Comparison: We perform a statistical comparison of rankings between *manual-image* and *random-image* corpus for ambiance and physical environment dimensions. We performed the Wilcoxon signed-rank statistical test [32], which is a non-parametric test to compare the mean ranks of two populations. At the 99% confidence level, we obtained a p -value $< 2.2 \times 10^{-16}$ for both dimensions, validating the hypothesis that manually selected images are more appropriate for describing ambiance and physical environment.

In summary, the results provide an answer to RQ1, in that manual selection of images from Foursquare places provide more suitable images to capture a place’s ambiance as they contain enough visual cues to gauge a place’s physical environment. Even though random image selection is a more scalable setting, these images are often not ideal to characterize the indoor environment, given the wider variability. Please note that in this section we report the summary statistics across all cities combined. Individual trends for each city are similar to the overall trends and are omitted for brevity.

IV. LABELING PLACE AMBIANCE

A priori, the ambiance of places is not known to zero-acquaintance visitors or observers. In this section, we address our second research question RQ2 i.e., whether reliable estimates of ambiance can be obtained using Foursquare images.

Based on the *manual-image* corpus of 900 images across 300 places, we design a second crowdsourcing experiment and asked crowd workers to rate indoor ambiance along 13 different physical and psychological dimensions where images served as stimuli to form place impressions.

A. Selection of Ambiance Categories

In order to select ambiance dimensions to characterize place ambiance, we base our methodology on prior work [7]. The authors proposed a rating instrument consisting of 41 dimensions for ambiance characterization. In our work, we chose 13 dimensions for which the corresponding intraclass correlations were amongst the highest as reported in [7]. Note that many dimensions in [7] did not reach enough inter-annotator agreement. We used a five-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5) to judge the ambiance labels, while [7] used a 3-point categorical scale (yes, maybe, no). Throughout the rest of the paper, we will use the terms dimensions and labels interchangeably in the context of ambiance categories. The list of selected labels is shown in alphabetical order in Table III.

B. Crowdsourcing Ambiance Impressions

To answer RQ2, crowdsourcing was employed to gather ambiance impressions. We used MTurk with the same worker qualification requirements as the first study. In each HIT, the workers were asked to view three images corresponding to a place, and then rate their personal impressions of the place ambiance based on what they saw. As part of the annotation interface, we ensured that workers viewed images in high-resolution (and not just the image thumbnails). People were given a previous definition of each ambiance category. Moreover, workers were not informed about the city under study to reduce potential bias and stereotyping associated to the city identity. We collected 10 annotations for each ambiance dimension across all 300 places. Consequently, we collected a total of 3,000 responses, for which every worker was reimbursed 0.15 USD per impression.

V. PLACE AMBIANCE LABELING RESULTS

In this section, we present the results of our second crowdsourcing experiment.

A. Ambiance Annotations Quality

For a total number of 3,000 available HITs in this experiment, a typical worker completed an average of 56 HITs, with one worker completing 270 HIT assignments. When compared to the first experiment, similar results were obtained for HIT

Label	Barcelona	New York City	Paris	Seattle	Singapore	Mexico City	Combined	Graham [7]
Artsy	0.81	0.66	0.69	0.72	0.80	0.76	0.76	0.63
Bohemian	0.63	0.58	0.54	0.54	0.66	0.66	0.62	0.67
Conservative	0.67	0.77	0.78	0.67	0.70	0.85	0.76	0.77
Creepy	0.54	0.62	0.60	0.57	0.32	0.62	0.59	0.81
Dingy	0.74	0.81	0.59	0.69	0.67	0.81	0.74	0.74
Formal	0.76	0.93	0.93	0.89	0.89	0.90	0.91	0.82
Sophisticated	0.68	0.91	0.90	0.85	0.80	0.87	0.86	0.70
Loud	0.80	0.81	0.76	0.74	0.82	0.82	0.80	0.74
Old-fashioned	0.82	0.46	0.78	0.45	0.72	0.67	0.72	0.67
Off the beaten path	0.58	0.62	0.39	0.54	0.61	0.59	0.58	0.73
Romantic	0.38	0.84	0.86	0.80	0.83	0.86	0.82	0.63
Trendy	0.69	0.71	0.50	0.43	0.68	0.85	0.69	0.58
Up-scale	0.69	0.91	0.90	0.85	0.81	0.85	0.86	0.76

TABLE III: $ICC(1, k)$ scores of 13 ambiance dimensions for each city. $ICC(1, k)$ scores obtained in [7] is also shown in the last column for comparison. Cells marked in **bold** are **not** statistically significant at $p < 0.01$.

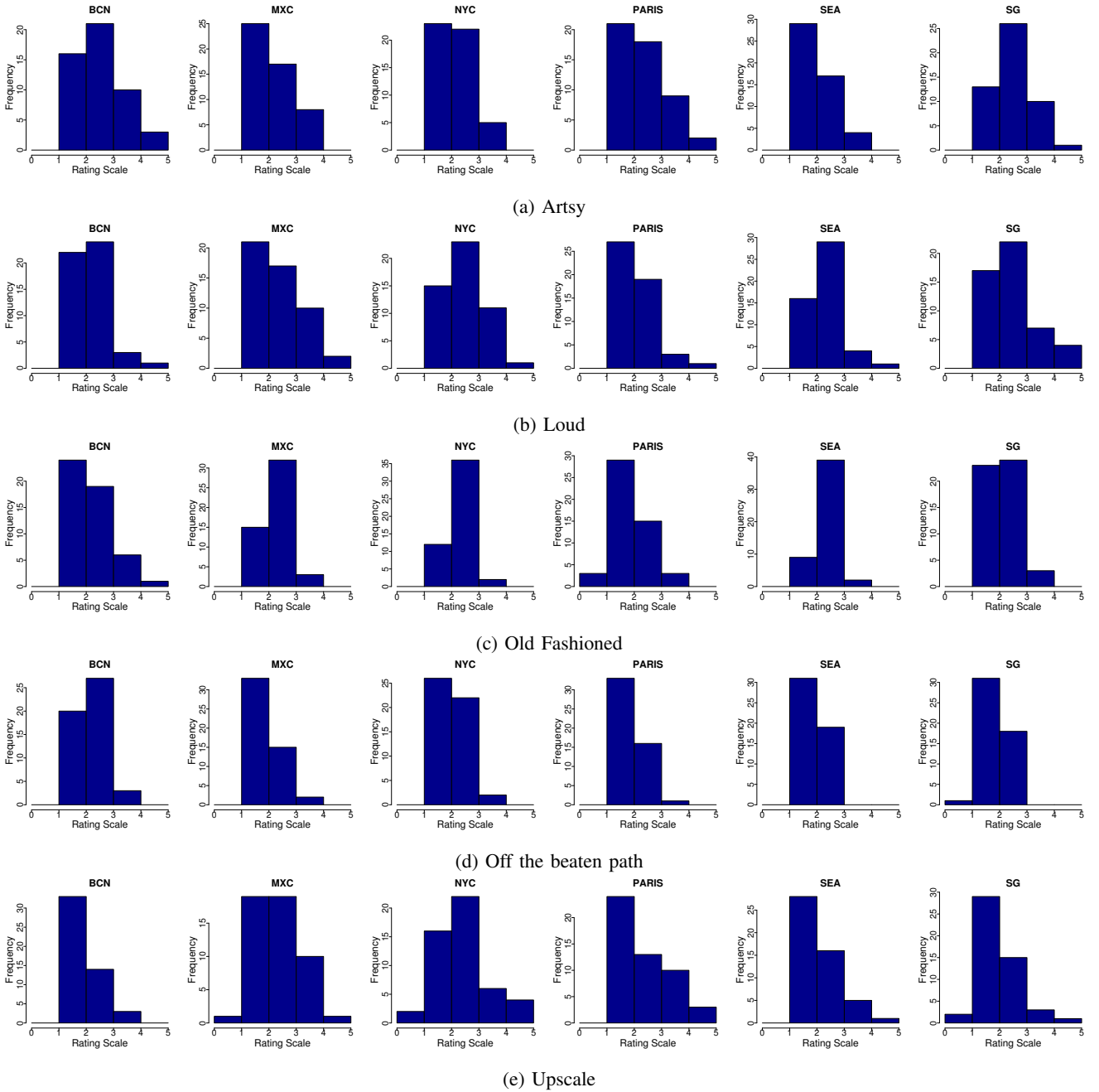


Fig. 5: Histograms showing the mean annotation scores across all cities for a) Artsy, b) Loud, c) Old fashioned, d) Off the beaten path, and e) Upscale.

Label	Barcelona	Mexico City	NYC	Paris	Seattle	Singapore	Combined
Artsy	2.54 (0.78)	2.20 (0.69)	2.14 (0.56)	2.36 (0.69)	2.05 (0.59)	2.46 (0.72)	2.29 (0.69)
Bohemian	2.34 (0.60)	1.94 (0.58)	2.07 (0.49)	2.09 (0.55)	1.99 (0.44)	2.04(0.57)	2.08 (0.55)
Conservative	2.04 (0.58)	2.36 (0.81)	2.33 (0.70)	2.17 (0.71)	2.37 (0.57)	2.28 (0.59)	2.26 (0.67)
Creepy	1.33 (0.31)	1.37 (0.38)	1.21 (0.27)	1.20 (0.27)	1.21 (0.27)	1.18 (0.19)	1.25 (0.29)
Dingy	1.68 (0.52)	1.61 (0.60)	1.60 (0.58)	1.49 (0.40)	1.57 (0.46)	1.49 (0.42)	1.57 (0.50)
Formal	1.60 (0.50)	2.13 (0.91)	2.14 (0.97)	2.01 (0.96)	1.95 (0.75)	1.62 (0.68)	1.91 (0.84)
Sophisticated	2.09 (0.56)	2.41 (0.90)	2.42 (0.96)	2.37 (0.93)	2.20 (0.74)	2.15 (0.70)	2.27 (0.82)
Loud	2.30 (0.67)	2.45 (0.78)	2.51 (0.72)	2.09 (0.68)	2.33 (0.62)	2.49 (0.83)	2.36 (0.73)
Old-fashioned	2.20 (0.77)	2.30 (0.62)	2.33 (0.46)	1.90 (0.67)	2.44 (0.41)	2.16 (0.56)	2.22 (0.61)
Off the beaten path	2.27 (0.51)	1.88 (0.53)	2.06 (0.52)	1.89 (0.46)	1.99 (0.45)	1.96 (0.48)	2.01 (0.50)
Romantic	1.77 (0.37)	2.09 (0.81)	1.95 (0.72)	1.92 (0.78)	1.86 (0.62)	1.80 (0.65)	1.90 (0.68)
Trendy	2.34 (0.65)	2.55 (0.89)	2.55 (0.66)	2.49 (0.55)	2.45 (0.47)	2.54 (0.64)	2.49 (0.65)
Up-scale	1.93 (0.56)	2.36 (0.85)	2.39 (0.93)	2.36 (0.91)	2.13 (0.70)	2.01 (0.69)	2.20 (0.80)

TABLE IV: Means and standard deviations (in brackets) of annotation scores for each city and label.

completion times (mean: 97 secs, median: 68 secs, max: 596 secs).

Now we turn our focus towards assessing the reliability of the annotations. We measure the inter-annotator consensus by computing intraclass correlation (ICC) among ratings given by the worker pool. Our annotation procedure requires every place to be judged by k annotators randomly selected from a larger population of K workers ($k = 10$, while K is unknown as we have no means to estimate the MTurk worker population). Consequently, $ICC(1, 1)$ and $ICC(1, k)$ values, which respectively stand for single and average ICC measures [33] are computed for each ambiance dimension across all cities.

Table III reports the $ICC(1, k)$ values for all cities (due to space constraints, we omit $ICC(1, 1)$ values.) In addition to listing the individual scores for each city and label, we also report the combined $ICC(1, k)$ scores for each label for the whole dataset, where we have combined all places across cities. We observe acceptable inter-rater reliability for most labels, with all the scores being statistically significant (p -value < 0.01), with the exception of *creepy* label in Singapore. We notice that the inter-rater reliability for labels *formal*, *sophisticated*, *romantic* and *up-scale* is typically high (above 0.8) for most of the cities. Using correlation analysis (presented in Section V-B2), we find that these four labels are collinear, with pairwise correlations exceeding 0.8. It is interesting to note that label *loud* achieved high agreement from images not showing any sound (0.8 combined score). On the other hand, labels *creepy* and *off the beaten path* are the labels with lowest ICC (below 0.6 for the combined score.)

Importantly, these reliability scores are comparable to the ones obtained by Graham et al. [7], who conducted a similar study, but where the raters physically visited every venue (see section II and last column of Table III), while in our case online images act as a stimuli. To summarize, these results provide an answer to RQ2 as they show that consistent impressions of place ambiance can be formed based upon images contributed in social media, which further suggests that there might be strong visual cues present for annotators to form accurate place impressions. The investigation of what specific cues contribute to impression formation will be the subject of future work.

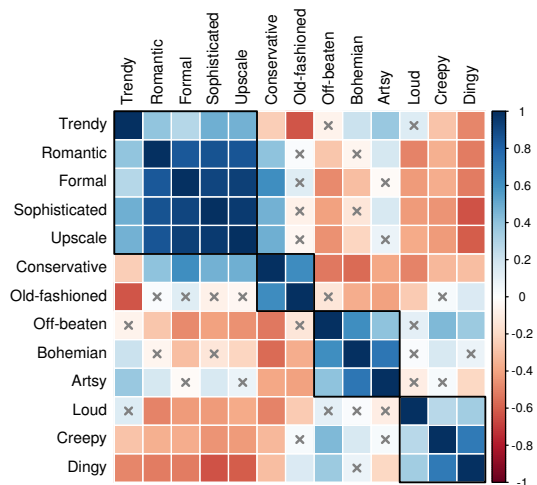


Fig. 6: Plot showing the correlation matrix between ambiance dimensions. Black rectangular borders indicate the four distinct clusters found in the correlation matrix. Cells marked **X** are **not** statistically significant at $p < 0.01$.

B. Comparing Impressions across Cities

In this section we present descriptive statistics and study differences across cities for each ambiance label.

1) *Descriptive Statistics*: Table IV lists the descriptive statistics (mean score and standard deviation) for each city and label. The mean scores are derived as follows: first, for every place we compute the mean score for each ambiance label, using the 10 annotations per label for each place; we then compute the mean scores and standard deviations for each city and label using all 50 places for each city. At the level of individual annotations, minimum and maximum values are 1 and 5 respectively for all each city and label, showing that the full scale is used by the crowd-workers. Note that the mean value obtained for all labels and all cities is below 3, which indicates a trend towards disagreement with the corresponding label. On the other hand, each city has venues that score high for each dimension.

In all cities, except Barcelona, the mean score for *trendy* is highest amongst all labels; Barcelona places score the maximum on being *artsy*. *Creepy* scores the lowest (along with the lowest variance) for all cities which is not surprising

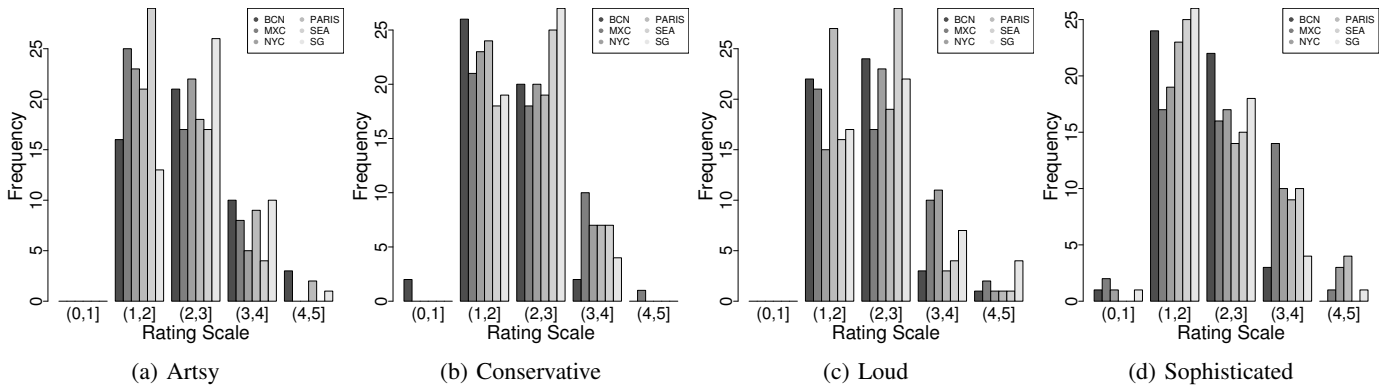


Fig. 7: Barplots comparing the mean annotation scores across all cities for a) Artsy, b) Conservative, c) Loud, and d) Sophisticated.

given that all places are popular places in their respective cities. From Table IV, we do not observe much variation in the mean values across cities, but a few differences stand out. For instance, the mean differences of the *formal* attribute between NYC and Barcelona, and the *old fashioned* attribute between Paris and Seattle exceed 0.5, potentially suggesting differences in place perceptions.

Figure 5 plots the distribution of mean annotation scores for five of the 13 labels across all cities. As observed in the previous paragraph, even though we do not observe much variation in mean values across cities (Table IV), comparing the distributions of some labels between cities provides additional insights to examine the difference. For example, the mean difference of *loud* attribute between NYC and Mexico City is small (difference of 0.06), but the respective distributions look significantly different, as shown in Figure 5b. The same is true for the *up-scale* attribute between NYC and Paris (difference of 0.03), as highlighted in Figure 5e. To visually aid the comparison between cities, we show the barplots of the binned mean annotation scores for four the ambiance labels in Figure 7, where finer differences can be observed across cities and relative ratings.

2) *Correlation Analysis*: To understand the statistical relationship between ambiance labels, we perform correlation analysis using mean annotation scores for all ambiance labels. Figure 6 visualizes the correlation matrix across all ambiance dimensions. We have used hierarchical clustering to re-order the correlation matrix in order to reveal its underlying structure. Note that we color code the matrix instead of providing numerical scores to facilitate the discussion. We observe four distinct clusters. Starting from top left in the first cluster, labels *formal*, *sophisticated*, *romantic* and *up-scale* are highly collinear with pairwise correlations exceeding 0.8. The second cluster consists of places which are either *conservative* or *old-fashioned*, and the third cluster consists of *off-beaten*, *bohemian* or *artsy* places. The fourth cluster (bottom-right) lies on the opposite spectrum with respect to cluster one, and consists of *loud*, *dingy* and *creepy* places. Each of these four clusters clearly correspond to different ambiances. Furthermore, we can also observe significant negative correlations between dimensions in cluster one and cluster four and between clusters two and tree.

3) *Statistical Comparison*: To better understand whether mean differences across cities for some of these ambiance labels are statistically significant, we perform the Tukey’s honest significant difference (HSD) test. Tukey’s HSD test is a statistical procedure for groups which compares all possible pairs of mean values for each group, with the null hypothesis stating that the mean values being compared are drawn from the same population [34]. We perform the HSD test to compute the pairwise comparisons of mean values between cities for each ambiance label, which result in a total of 195 comparisons (15 city-wise pairs across 13 dimensions). Table V lists only the significant results of the Tukey’s HSD test, where the differences in the observed means are statistically significant at p -value < 0.01 . In addition, we plot some of the significant results from Tukey’s HSD for two city pairs in Figure 8. Based on these statistics, we observe that:

- 1) Popular places in Barcelona are perceived as more *bohemian* and *artsy* compared to places in Mexico City and Seattle respectively (see Figure 7 and Figure 8a);
- 2) Popular places in Paris are perceived as less *old fashioned* compared to New York City (Figure 8b) and Seattle; and,
- 3) Barcelona places are perceived to be more *off-beaten* when compared to places in both Mexico City and Paris.

To validate the statistical significance of the Tukey’s HSD test, we perform a series of pairwise Kolmogorov-Smirnov test (KS test) across all cities and labels. The KS test is a non-parametric test to compare the empirical distributions of two samples, with the null hypothesis being that the two samples are drawn from the same distribution [35]. We perform the KS test to compare the cumulative distribution functions of each city-pair across each dimension (195 comparisons). We report the p -values for the KS test in Table V for a statistical level $\alpha = 0.01$. Results from the KS test confirms most of the findings from the Tukey’s HSD test. It is worth noting that if we increase the significance level (α) to 0.05, we observe differences across other city-pairs to be statistically significant as well, but we have not listed them due to space restrictions.

To conclude this subsection, our study shows that most of the differences across cities for each of the ambiance dimensions are not statistically significant. This result is interesting in itself as it might suggest that popular places in social media in cosmopolitan cities have many points in common. To our

Label	City Pair	Mean Difference	p -value $\times 10^{-3}$
Bohemian	MEX-BCN	-0.398	3.70 (39.68)
Artsy	SEA-BCN	-0.492	4.26 (6.18)
Old Fashioned	PAR-NYC	-0.434	4.09 (0.67)
Old Fashioned	SEA-PAR	+0.544	9.9×10^{-2} (7.4×10^{-3})
Off the beaten path	MEX-BCN	-0.386	1.43 (0.051)
Off the beaten path	PAR-BCN	-0.376	2.11 (0.67)

TABLE V: Tukey’s HSD and KS test statistics. p -values obtained from KS test are shown in brackets in the last column. Values marked in **bold** are **not** statistically significant at $p < 0.01$

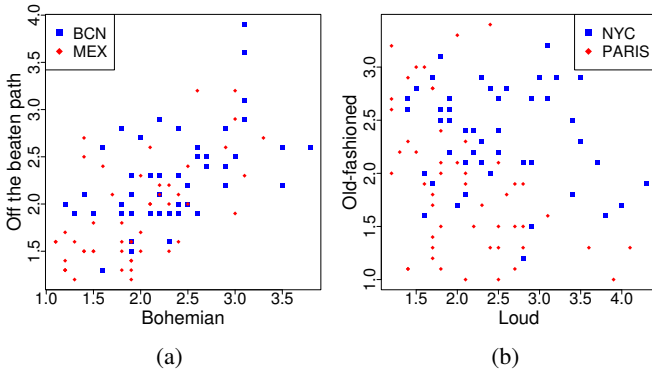


Fig. 8: Plot showing the comparison of a) *bohemian* and *off the beaten path* attribute between BCN and MEX, and b) *old fashioned* and *loud* attribute between NYC and Paris.

knowledge, this is the results that has not been reported before in social media research, but that could have some support from literature that talks about the “uniformization of taste” in globalized cities and social media content. This said, any possible interpretation would have to be further validated e.g., by a combination of further data analysis and ethnography. In addition, these results highlight the need to study *other* venues, including not so popular places on Foursquare and places not represented on Foursquare because of well-known socio-economic biases in social media. We also plan to investigate this issue as part of future work.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a new study on how images collected from social media sites like Foursquare relate to the characterization of indoor ambiance in popular urban places. Using more than 50,000 images collected from 300 popular indoor places across six cities, we first assessed the suitability of social images as data source to convey place ambiance, and found through a crowdsourcing experiment that images with clear views of the environment are more informative of ambiance than other image categories. Second, we demonstrated through another crowdsourcing experiment that reliable estimates of ambiance for several dimensions can be obtained using Foursquare images, suggesting the presence of strong visual cues to form accurate place impressions. Furthermore, we found that most aggregated impressions of popular places are similar across cities, but few statistically

significant differences across four ambiance dimensions (*bohemian*, *artsy*, *old fashioned*, and *off the beaten path*) exist between cities.

Our work contributes to multimedia research through the study of a largely unexplored research problem with both scientific and practical value. We will make the dataset publicly available after the publication of the paper. For future work, we first want to understand what specific cues in the indoor physical environment people use to judge a place and form ambiance impressions – whether its color, lighting scheme, spatial layout, or interior design – when looking at social media content. Understanding ambiance is informative for place owners as it can help them understand and potentially change the perception the place evokes. Second, we can think of ambiance as an additional place search feature in online venue platforms. For instance, in addition to let users search a place by its function or type of cuisine it offers, ambiance information would let users search a place by its ambiance type e.g., a *formal* place for a family dinner, or a *romantic* place for a date. This could complement existing sources of information like place reviews. We plan to pursue these two research directions.

REFERENCES

- [1] E. Glaeser, *Triumph of the city: How our greatest invention makes us richer, smarter, greener, healthier and happier*. Macmillan, 2011.
- [2] S. Zukin, *The cultures of cities*. Blackwell Oxford, 1995.
- [3] S. Carr, *Public space*. Cambridge Press, 1992.
- [4] R. Florida, P. Rentfrow, K. Sheldon, T. Kashdan, and M. Steger, “Place and well-being,” *Designing positive psychology: Taking stock and moving forward*, 2011.
- [5] P. J. Rentfrow, “The open city,” *Handbook of Creative Cities*, p. 117, 2011.
- [6] S. D. Gosling, S. Gaddis, and S. Vazire, “First impressions based on the environments we create and inhabit,” *First impressions*, 2008.
- [7] L. Graham and S. Gosling, “Can the ambiance of a place be determined by the user profiles of the people who visit it,” in *Proc. AAAI ICWSM*, 2011.
- [8] S. Bakhshi, P. Kanuparth, and E. Gilbert, “Demographics, weather and online reviews: A study of restaurant recommendations,” in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW ’14. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014, pp. 443–454. [Online]. Available: <http://dx.doi.org/10.1145/2566486.2568021>
- [9] A. Kittur, E. H. Chi, and B. Suh, “Crowdsourcing user studies with mechanical turk,” in *Proc. CHI*. ACM, 2008, pp. 453–456.
- [10] D. H. Kim, J. Hightower, R. Govindan, and D. Estrin, “Discovering semantically meaningful places from pervasive rf-beacons,” in *Proceedings of the 11th international conference on Ubiquitous computing*. ACM, 2009, pp. 21–30.
- [11] A. Quattoni and A. Torralba, “Recognizing indoor scenes,” in *CVPR*. IEEE, 2009.
- [12] H. Wang, D. Lymberopoulos, and J. Liu, “Local business ambiance characterization through mobile audio sensing,” in *Proc. WWW*, 2014, pp. 293–304. [Online]. Available: <http://dx.doi.org/10.1145/2566486.2568027>
- [13] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, “Soundsense: scalable sound sensing for people-centric applications on mobile phones,” in *Proc. MobiSys*. ACM, 2009.
- [14] D. H. Kim, K. Han, and D. Estrin, “Employing user feedback for semantic location services,” in *Proc. ACM UbiComp*. ACM, 2011, pp. 217–226.
- [15] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao, “Automatically characterizing places with opportunistic crowdsensing using smartphones,” in *Proc. UbiComp*. ACM, 2012, pp. 481–490.
- [16] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz, “On the semantic annotation of places in location-based social networks,” in *Proc. KDD*. ACM, 2011, pp. 520–528.

- [17] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history," *ACM Transactions on the Web (TWEB)*, vol. 5, no. 1, p. 5, 2011.
- [18] Y. Chen, A. Cheng, and W. Hsu, "Travel recommendation by mining people attributes and travel group types from community-contributed photos," 2013.
- [19] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes paris look like paris?" *ACM Trans. Graph.*, vol. 31, no. 4, p. 101, 2012.
- [20] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 761–770.
- [21] Y. Hu, L. Manikonda, and S. Kambhampati, "What we instagram: A first analysis of instagram photo content and user types," in *Proc. AAAI ICWSM*, 2014.
- [22] S. Bakhshi, D. A. Shamma, and E. Gilbert, "Faces engage us: photos with faces attract more likes and comments on instagram," in *Proc. ACM CHI*. ACM, 2014, pp. 965–974.
- [23] J. Baker, D. Grewal, and A. Parasuraman, "The influence of store environment on quality inferences and store image," *Journal of the Academy of Marketing Science*, vol. 22, no. 4, pp. 328–339, 1994.
- [24] C. C. Countryman and S. Jang, "The effects of atmospheric elements on customer impression: the case of hotel lobbies," *International Journal of Contemporary Hospitality Management*, vol. 18, no. 7, pp. 534–545, 2006.
- [25] L. W. Turley and R. E. Milliman, "Atmospheric effects on shopping behavior: a review of the experimental evidence," *Journal of Business Research*, vol. 49, no. 2, 2000.
- [26] V. Heung and T. Gu, "Influence of restaurant atmospherics on patron satisfaction and behavioral intentions," *International Journal of Hospitality Management*, vol. 31, no. 4, pp. 1167–1177, 2012.
- [27] Y. Liu and S. S. Jang, "The effects of dining atmospherics: an extended mehrabian–russell model," *International Journal of Hospitality Management*, 2009.
- [28] P. Salesses, K. Schechtner, and C. A. Hidalgo, "The collaborative image of the city: Mapping the inequality of urban perception," *PLoS ONE*, 2013. [Online]. Available: <http://dx.doi.org/10.1371/journal.pone.0068400>
- [29] D. Quercia, N. K. O'Hare, and H. Cramer, "Aesthetic capital: what makes london look beautiful, quiet, and happy?" in *Proc. CSCW*. ACM, 2014.
- [30] J. Frith, "Constructing location, one check-in at a time: Examining the practices of foursquare users," Ph.D. dissertation, North Carolina State University, 2012.
- [31] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: Shifting demographics in mechanical turk," in *Proceedings ACM CHI '10 Extended Abstracts*, 2010. [Online]. Available: <http://doi.acm.org/10.1145/1753846.1753873>
- [32] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, no. 6, pp. 80–83, 1945.
- [33] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.
- [34] J. W. Tukey, "The philosophy of multiple comparisons," *Statistical science*, pp. 100–116, 1991.
- [35] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, 1951.