InnerView: Learning Place Ambiance from Social Media Images

Darshan Santani
Idiap Research Institute
EPFL, Lausanne, Switzerland
dsantani@idiap.ch

Rui Hu
Idiap Research Institute
rhu@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
EPFL, Lausanne, Switzerland
gatica@idiap.ch

ABSTRACT

In the recent past, there has been interest in characterizing the physical and social ambiance of urban spaces to understand how people perceive and form impressions of these environments based on physical and psychological constructs. Building on our earlier work on characterizing ambiance of indoor places, we present a methodology to automatically infer impressions of place ambiance, using generic deep learning features extracted from images publicly shared on Foursquare. We base our methodology on a corpus of 45,000 images from 300 popular places in six cities on Foursquare. Our results indicate the feasibility to automatically infer place ambiance with a maximum R^2 of 0.53 using features extracted from a pre-trained convolutional neural network. We found that features extracted from deep learning with convolutional nets consistently outperformed individual and combinations of several low-level image features (including Color, GIST, HOG and LBP) to infer all the studied 13 ambiance dimensions. Our work constitutes a first study to automatically infer ambiance impressions of indoor places from deep features learned from images shared on social media.

1. INTRODUCTION

There is increasing interest in characterizing the physical and social ambiance of urban spaces (both indoor and outdoor), to understand how people perceive these environments based on physical and psychological constructs [7, 21, 22, 23]. Ambiance, formally defined as the "character and atmosphere of a place" [4], is the human way of relating to places. Consequently, understanding ambiance of places in cities has been proposed as a new topic in multimedia research [22]. Characterizing place ambiance has many applications ranging from hyper-local, ambiance-driven place search and discovery (e.g., a *trendy* place for a night-out or a *romantic* place for the wedding anniversary) to data-driven recommendations for place owners to improve the presentation (e.g., architecture design and style) of their venues.

In hospitality research, there is a significant body of work that investigates the effect of ambiance on patrons' dining experience, perception, and customer retention [1, 25, 8], but most of these studies are small-scale and based on in-situ interviews and questionnaires, which may have limitations with respect to generalization or recall biases. To overcome these limitations, researchers have used geo-tagged images from Google Street View or taken by volunteers to measure the perceptions of outdoor spaces for several urban dimensions including safety, quietness, etc. [21, 16, 19, 23].

Until recently, most of the studies examining urban perceptions using online images were focused on outdoor spaces. From the perspective of urban design, studying elicited impressions for indoor places should involve examination of different variables when compared to outdoor spaces; for instance, "formal" is a meaningful ambiance construct in an indoor venue, but not necessarily so in an outdoor space. As a brave new topic in [22], we presented a crowdsourcing methodology to examine the suitability of social media images for the characterization of indoor ambiance impressions across 13 dimensions (including artsy, romantic, formal, loud and trendy, among others) in 300 popular Foursquare places. We found that reliable estimates of ambiance were obtained for several of the dimensions, suggesting the presence of visual cues that allow to create such impressions. In this paper, we extend our earlier work to automatically characterize place ambiance using visual cues from images.

To infer human perception of outdoor spaces, recent works have used a variety of low-level image features including color, GIST, HOG, LBP, SIFT and generic deep convolutional activation features. Using these features on geo-tagged images from Google Street View, high-level attributes (e.g., wealthy, uniqueness, and safety) for outdoor scenes are inferred in two US cities [11, 14]. Using the same dataset, in [15], a CNN architecture is proposed to predict and discover mid-level visual patterns which correlate with the perceived safety of an outdoor scene. Building upon these works, in this paper we focus on studying places across 13 dimensions appropriate for the indoor environment. Specifically, we address two research questions:

RQ1: Can judgments of ambiance of an indoor place be automatically inferred using low-level image and generic deep features extracted from social media images?

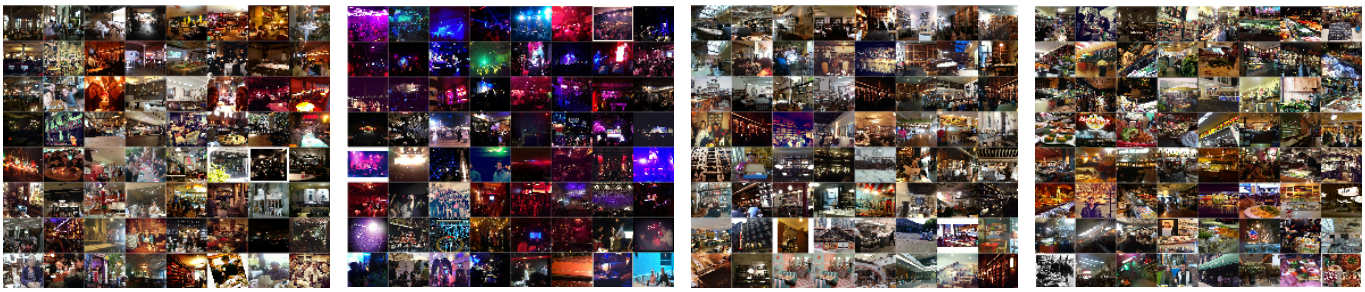
RQ2: What visual categories represent popular places on Foursquare in connection to indoor ambiance?

To address these questions, we build upon our earlier work and dataset [22]. For automatic characterization and inference, we build upon prior work in object classification and scene understanding using deep learning techniques [20, 17, 26]. Our contributions are two-fold. First, we devised a methodology to automatically infer ambiance impressions of popular Foursquare places, using generic deep features extracted from images. We base our methodology using a corpus of over 45,000 images from 300 popular places on

MM '16, October 15 - 19, 2016, Amsterdam, Netherlands

ISBN 978-1-4503-3603-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2964284.2967261



(a) Restaurant, Eating Place (b) Stage (c) Library (d) Grocery Store

Figure 1: Sample of random images from the *50K image* corpus which were classified as (a) Restaurants/Eating Place, (b) Stage, (c) Library, and (d) Grocery Store. For each class, from top to bottom, images are sorted in decreasing order of ImageNet dominant class probability (Section 4.1).

Foursquare spread across six cities. The results indicate the feasibility to automatically infer the ambiance of these places with a maximum R^2 of 0.53 using features extracted from a pre-trained convolutional neural network, precisely GoogLeNet model trained on ImageNet data. Furthermore, we found that CNN features consistently outperform individual and combinations of several low-level image features for all the 13 studied dimensions. Second, to better understand the image corpus, we studied the distribution of the most likely ImageNet class assigned to each image. We find that most of the top-10 dominant classes are associated with either food or drinks, but also contains images which represent the physical environment with clear and unoccluded views of the indoor space. These findings point towards the nature of images shared on Foursquare, which to our knowledge have not been analyzed before at such a scale from the perspective of ambiance [7].

2. DATASET

We use the image corpora and place ambiance annotations collected as part of our prior work [22], which contains data from 300 popular places on Foursquare in six cities – Barcelona, Mexico City, New York City, Paris, Seattle, and Singapore. For each city, we chose 50 popular places among restaurants, cafes, bars, or nightclubs. Below we describe the data:

Image Corpora: For each selected place, there are two image corpora: a) *Physical Environment Image Corpus*: this corpus was manually curated to contain images with clear views of the physical environment of a place. It contains three images per place, which show the indoor space from different angles, resulting in a total of 900 images for all 300 places; b) *50K Image Corpus*: the second image corpus contains all publicly available images shared on Foursquare for each place, resulting in a total of 45,848 images for 280 places (an average of 164 images per place). Note that the 50K image corpus contains images for 280 places, as opposed to 300 places, due to changes in the Foursquare API [6]. Figure 1 shows a sample of images selected randomly across four ImageNet categories from this corpus.

Ambiance Annotations: In addition to images, the data contains the results of an online crowdsourcing study to collect ambiance impressions for each place, based on *physical environment image* corpus. Ambiance ratings were elicited across 13 physical and psychological dimensions (shown in Table 1), where images served as stimuli to form place impressions. Ambiance ratings were obtained using the three manually selected images for each place. For this study, we used the mean annotation ratings for each place and dimension as described in [22].

3. METHODOLOGY

3.1 Feature Extraction

Building upon recent work in the literature, we have extracted the following set of low-level and deep visual features:

1. Color: We computed global color histogram in RGB space. Each channel was quantized into 8-bins, resulting in 8^3 possible color combinations and a 512-dimensional color feature vector.

2. GIST: This descriptor captures the dominant spatial structure of a scene from a set of perceptual dimensions (e.g., naturalness, openness, roughness, etc.) [13]. We use the standard setting of this descriptor, resulting in a 512-dimensional vector.

3. Texture (LBP): Texture captures the spatial arrangement of color and intensities in an image. We apply the local binary pattern (LBP) descriptor [12], which encodes local texture information (such as spots, edges, and corners) by comparing each pixel with its neighborhood pixels, resulting in a 256-dimensional vector.

4. Gradient (HOG): Histogram of oriented gradients (HOG) computes occurrences of gradient orientations in localized region of an image [10]. We apply the pyramid HOG implementation, where images are first represented in pyramid hierarchies, then the HOG descriptor is computed on each level, and finally the final descriptor is the concatenation of vectors across all levels [2]. We compute the pyramid HOG descriptor for levels $l = 0$ to $l = 3$. Images are divided into 2^{2*l} regions, and a 8-bin histogram is computed within each region, which results in a 680-dimensional feature vector.

5. CNN: The availability of large-scale image datasets [20] and the performance of deep neural networks for object classification and scene understanding [17, 26], have opened opportunities to explore these features for our problem. We have used the features extracted using a pre-trained convolutional neural networks model (CNNs) using the Caffe framework [9]. Specifically, we used the GoogLeNet CNN [24] trained on ImageNet data. ImageNet data contains over 14 million images across 1,000 categories. To extract the CNN descriptors, for each image, we obtained the final layer class probabilities across all 1,000 ImageNet classes, resulting in a 1000-dimensional feature vector.

We chose to use a pre-trained model trained on a large and diverse data (ImageNet in our case), as opposed to training a CNN model on our data for two reasons. First, it has been shown that features extracted using pre-trained CNN models can potentially provide discriminative features for multiple visual recognition tasks [5, 17]. ImageNet categories are well suited for our problem as they are descriptive of visual cues typically present in restaurants, bars, etc. (refer to Figures 1 and 2). Second, using a pre-trained model avoids the need to train, adapt or fine-tune a CNN on our dataset, which can be computationally expensive and resource intensive.

Feature Aggregation: As stated in Section 2, ambiance ratings were given for each place and each place had an average of 164 images. We extracted all the previously described visual features for each image. Then, in order to obtain a representative feature vector for each place, we apply an early feature fusion approach by computing the mean feature vectors of all images describing the same place, for each feature set.

3.2 Inference Method and Evaluation

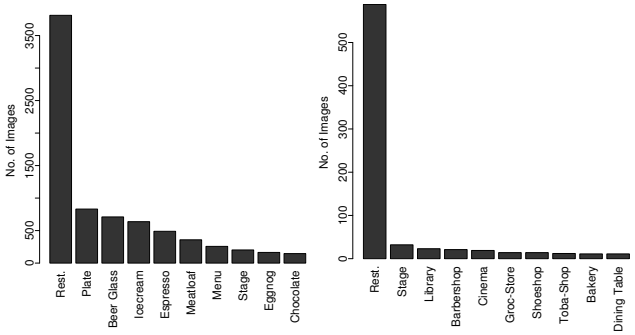
We are interested in examining the predictive power of both the low-level image features and deep CNN features to automatically infer the perceptions of social ambiance for the studied dimensions. We approach this problem as a regression task, where the objective is to infer the mean annotation scores of each place. For regression, we use Random Forest, a tree-based supervised learning method that guards against overfitting to the training data [3]. In all our experiments we used a 10-fold cross-validation approach. To evaluate the performance of different feature sets, we used two standard measures: the root-mean-square error ($RMSE$) and coefficient of determination (R^2) between the perceived ground-truth and inferred ambiance scores for each label and feature set. Furthermore, we have used variable importance measures from random forests to understand the relative importance of visual cues for each dimension (see Section 4). To understand and compare the predictive performance of each feature set, we choose the baseline model to be the mean annotated score as the predicted value for each label.

4. RESULTS AND DISCUSSION

4.1 Visual Categories

We begin our analysis by examining the distribution of the most likely ImageNet class assigned to each image. As stated before, the last layer of the GoogLeNet CNN model outputs the probability distribution of the image across all 1,000 ImageNet classes. Given this probability distribution, we chose the ImageNet class with the highest probability as the dominant class for each image. In Figure 2, we show the distribution of the top-10 dominant classes for both image corpora. For the *50K image* corpus, most of the top ten dominant categories are associated with either food (e.g., plate, meatloaf, ice-cream, chocolate) or drinks (e.g., beer glass, espresso, eggnog), as shown in Figure 2a. These results are consistent with previous findings [22] and expected given that all places are restaurants, bars, cafes, or nightclubs (Section 2).

While analyzing the top-10 class distribution for the *physical environment image* corpus, we observe that most of the dominant classes relate to the physical attributes of the indoor environment (e.g., stage, library, dining table, bakery, etc.) and do not contain food or drinks categories, which is in contrast with the class distribution for the *50K image* corpus (Figure 2b). These findings are not surprising given that all images in this corpus were manually selected to show clear views of the indoor scene (Section 2) [22]. Some of the recognized categories in *physical environment image* corpus may seem intriguing at first glance (e.g., library, grocery store, or barbershop), but after manually browsing the images belonging to these categories, we found that these classes describe various attributes of the indoor environment and are misclassified yet makes sense visually. For instance, most of the images belonging to "library" class contain images showing wall shelves typically found in cafes and bars (Figure 1c); while some of the "grocery store" images contain transparent window shelves displaying food or drink items (Figure 1d).



(a) 50K Image Corpus (b) Phy. Env. Image Corpus

Figure 2: Histogram of Top 10 recognized ImageNet classes for a) *50K Image* Corpus, and b) *Physical Environment Image* Corpus

Further, we observe that the top class in both image corpora is "restaurant/eating place". For the *50K* and *physical environment* image corpora, this class respectively contains 8% and 65% of total images. At first, it looks like a generic class but after manually browsing these images (Figure 1a), we found that most of the images represent the physical environment well, providing clear and unoccluded views of the indoor space. These findings point towards the feasibility to automate the selection of images for indoor ambiance characterization. As we previously reported, human observers preferred views of the physical environment of places to make impressions of ambiance [22]. However, as we show in the following subsection, other image categories are also exploited by the learning algorithm.

While the plots shown in Figure 2 are focused on illustrating the dominant visual class, many other classes are also informative of ambiance for a given image (e.g., a photo of someone eating an ice cream in a restaurant or drinking an espresso in a coffee shop are not likely to be generated in a club). This is the reason why we use the 1000-dimensional probability vector to infer the ambiance dimensions, as described in the next subsection. Overall, these findings point towards the nature of images shared on Foursquare, which to our knowledge has not been analyzed before at such a scale from the perspective of ambiance.

4.2 Regression

In this section, we evaluate the performance of both low-level image features and deep CNN features to automatically infer impressions of place ambiance. In Table 1, we report the R^2 values between ground-truth and inferred ambiance scores for each label and feature set. For the baseline model and the model learned using CNN features, we also report the $RMSE$ values. All the low-level visual features (color, GIST, HOG, and LBP) are computed on the *50K image* corpus, while the CNN features are computed for both image corpora.

Overall, the results indicate that a maximum R^2 of 0.53 can be obtained by the CNN-based regressor (for the *loud* dimension), while low R^2 values are obtained for other dimensions (*creepy*, *dingy*, *off the beaten path*). For six out of 13 dimensions, the obtained R^2 exceeds 0.30. On the *50K image* corpus, we find that CNN features consistently outperform individual and combinations of several low-level image features (including color, GIST, HOG, and LBP) for all labels, consistent with results reported in the vision and multimedia literature [17]. Note that in Table 1, we have not shown the results for the combination of low-level image features due to space constraints. The results highlight the suitability of pre-trained CNN models to infer perceptual judgments in indoor

	Baseline-50K		Color-50K	GIST-50K	HOG-50K	LBP-50K	CNN-Phy-Env		CNN-50K	
	R^2	$RMSE$	R^2	R^2	R^2	R^2	R^2	$RMSE$	R^2	$RMSE$
Artsy	0.0	0.69	0.01	0.02	0.04	0.05	0.12	0.66	**0.22**	0.63
Bohemian	0.0	0.55	0.08	0.05	0.11	0.09	0.08	0.54	**0.24**	0.50
Conservative	0.0	0.67	0.21	0.20	0.19	0.11	0.24	0.60	**0.30**	0.57
Creepy	0.0	0.29	0.05	0.04	0.01	0.00	0.06	0.29	**0.14**	0.28
Dingy	0.0	0.50	0.04	0.02	0.01	0.05	0.05	0.50	**0.17**	0.47
Formal	0.0	0.82	0.10	0.07	0.03	0.10	0.28	0.72	**0.37**	0.70
Loud	0.0	0.73	0.33	0.29	0.26	0.31	**0.53**	0.51	0.52	0.51
Off the beaten path	0.0	0.61	0.05	0.01	0.01	0.00	0.15	0.47	**0.17**	0.47
Old-fashioned	0.0	0.50	0.16	0.11	0.10	0.08	**0.24**	0.54	0.22	0.55
Romantic	0.0	0.67	0.10	0.15	0.03	0.08	0.36	0.57	**0.39**	0.56
Sophisticated	0.0	0.79	0.11	0.10	0.04	0.10	0.26	0.72	**0.38**	0.67
Trendy	0.0	0.64	0.19	0.1	0.12	0.15	0.17	0.61	**0.32**	0.54
Up-scale	0.0	0.78	0.14	0.11	0.03	0.13	0.29	0.69	**0.40**	0.65

Table 1: Inference results for 13 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 across all feature sets.

places; our finding complements what has been shown in recent literature regarding deep learning of high-level perception of outdoor scenes [14, 15].

While evaluating the performance of deep features between the *50K* and *physical environment* image corpora, we observe that the CNN-50K model outperforms the CNN-Phy-Env model in terms of higher R^2 and lower $RMSE$ values for most of the dimensions except the *loud* and *old-fashioned* dimensions. When examining the differences in more detail, we find that for some labels (e.g., *loud*, *romantic*, *old-fashioned*, *off the beaten path*), R^2 values are comparable, while for some dimensions (e.g., *artsy*, *bohemian*, *trendy*), the difference between R^2 values across image corpora is relatively high. Note that the ambiance ratings were obtained using just three manually selected images. These findings suggest that it is difficult to automatically represent some dimensions with just three images (compared to the human ability to infer and generalize from very few examples), and that the availability of more images provide additional informative visual cues to characterize some of these labels at high level. Future work will investigate the effect of the amount of data used for learning ambiance.

The performance of low-level features indicates that all feature sets perform the best for the *loud* feature (Table 1). Amongst the low-level feature sets, color features achieve the highest R^2 for most of the ambiance dimensions, including *trendy* and *up-scale* places, which corroborates results reported in [18], that attempted to infer ambiance for a 50-place dataset from profile pictures of Foursquare users rather than using photos of the venues. HOG and GIST also perform moderately for the *conservative* label.

While examining individual labels, the *loud* label achieves the highest performance with $R^2 \geq 0.26$ for all feature sets, suggesting visual patterns such as texture cues associated with the perception of loudness e.g., presence of crowd, an elevate stage, etc. (see Figure 1b). On the other hand, *creepy* achieves the lowest predictive performance ($R^2 \leq 0.14$). The low R^2 values for *creepy* can be attributed to a combination of lower inter-rater agreement and lowest mean annotation scores across all labels [22]. Overall, positively phrased (*romantic*, *up-scale*, *sophisticated*) and negatively phrased labels (*creepy*, *dingy*), which likely correspond to different ambiances, achieve in each case similar R^2 values.

Moreover, the *romantic* label achieves promising prediction performance ($R^2 = 0.39$) with CNN features, and relatively poor performance with low-level image features. We further analyze the relative importance of visual categories for different labels using the variable importance measures taken from random forests [3]. Using these measures, we find that "dining table" and "table lamp" are the top two discriminative visual categories for *romantic* places (see Figure 3a and 3b); "suit of clothes" and "red wine" for places perceived as *upscale* and *formal* (see Figure 3c); and "altar" for *artsy* places (see Figure 3d). In summary, the regression results suggest the feasibility to automatically infer place ambiance with promising prediction performance for some of the dimensions.

(a) Dining Table (b) Table Lamp (c) Suits (d) Altar

Figure 3: Sample of images from the *50K image* corpus which are recognized to belong to the visual category of a) Dining Table, b) Table Lamp, c) Suits of clothes, and d) Altar. For privacy reasons, images showing faces have been pixelated.

5. CONCLUSIONS

In this paper, we presented a methodology to automatically infer human impressions of place ambiance from social media images. Our results demonstrated the feasibility to automatically infer place ambiance using visual cues extracted from a pre-trained CNN model. We found that CNN features consistently outperformed low-level image features. Our work constitutes a first study to automatically characterize ambiance impressions of popular indoor places from deep features learned from social media images. Future work includes understanding the specific combination of different image types (food, people, environment) to ambiance inference, as well as the effect of the amount of available data.

6. ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation through the Youth@Night project. We thank Laurent Nguyen (Idiap) for discussions.

7. REFERENCES

[1] Julie Baker, Dhruv Grewal, and Ananthanarayanan Parasuraman. 1994. The Influence of Store Environment on Quality Inferences and Store Image. *Journal of the Academy of Marketing Science* 22, 4 (1994), 328–339.

[2] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Representing Shape with a Spatial Pyramid Kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR '07)*. ACM, New York, NY, USA, 401–408.

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

[4] Oxford Dictionary. 2015. Definition of Ambiance in English. http://bit.ly/1dAy80r. (2015). Retrieved July 24, 2016.

[5] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *ICML*. 647–655.

[6] Foursquare. 2015. Foursquare API August 2014 Update. https://developer.foursquare.com/docs/2014update. (2015). Retrieved July 24, 2016.

[7] L Graham and S Gosling. 2011. Can the Ambiance of a Place be Determined by the User Profiles of the People Who Visit It?. In *Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM)*.

[8] Vincent Heung and Tianming Gu. 2012. Influence of Restaurant Atmospherics on Patron Satisfaction and Behavioral Intentions. *International Journal of Hospitality Management* 31, 4 (2012), 1167–1177.

[9] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 675–678.

[10] Frederic Jurie and Bill Triggs. 2005. Creating Efficient Codebooks for Visual Recognition. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV '05)*. IEEE Computer Society, Washington, DC, USA, 604–610.

[11] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo. 2014. Streetscore – Predicting the Perceived Safety of One Million Streetscapes. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '14)*. IEEE Computer Society, Washington, DC, USA, 793–799.

[12] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 7 (July 2002), 971–987.

[13] 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.

[14] Vicente Ordonez and Tamara L Berg. 2014. Learning High-Level Judgments of Urban Perception. In *European Conference on Computer Vision*. Springer, 494–510.

[15] 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 Annual ACM Conference on Multimedia Conference*. ACM, 139–148.

[16] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. 2014. Aesthetic Capital: What makes London Look Beautiful, Quiet, and Happy?. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM.

[17] Ali Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features off-the-shelf: an Astounding Baseline for Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 806–813.

[18] Miriam Redi, Daniele Quercia, Lindsay Graham, and Samuel Gosling. 2015. Like Partying? Your Face Says It All. Predicting the Ambiance of Places with Profile Pictures. In *Ninth International AAAI Conference on Web and Social Media*.

[19] 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 '14)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 74–79.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[21] 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).

[22] 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 Annual ACM Conference on Multimedia Conference*. ACM, 211–220.

[23] 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.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[25] Lou W Turley and Ronald E Milliman. 2000. Atmospheric Effects on Shopping Behavior: A Review of the Experimental Evidence. *Journal of Business Research* 49, 2 (2000), 193–211.

[26] 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.