Ambiance in Social Media Venues: Visual Cue Interpretation by Machines and Crowds

Gülcan Can Yassir Benkhedda Daniel Gatica-Perez
Idiap Research Institute and École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
gulcan.can, yassir.benkhedda, gatica@idiap.ch

Abstract

We study the perception of ambiance of places captured in social media images by both machines and crowdworkers.This task is challenging due to the subjective nature of the ambiance construct as well as the large variety in layout, style, and visual characteristics of venues. For machine recognition of ambiance, we use state-of-the-art Residual Deep Convolutional Neural Networks (ResNets), followed by gradient-weighted class activation mapping (Grad-CAM) visualizations. This form of visual explanation obtained from the trained ResNet-50 models were assessed by crowdworkers based on a carefully designed crowdsourcing task, in which both visual ambiance cues of venues and subjective assessment of Grad-CAM results were collected and analyzed. The results show that paintings, photos, and decorative items are strong cues for artsy ambiance, whereas type of utensils, type of lamps and presence of flowers may indicate formal ambiance. Layout and design-related cues such as type of chairs, type of tables/tablecloth and type of windows are noted to have impact for both ambiances. Overall, the ambiance visual cue recognition results are promising, and the crowd-based assessment approach may motivate other studies on subjective perception of place attributes.

1. Introduction

Recently, image understanding has advanced considerably for many tasks [18], thanks to large-scale crowdsourced data collections, e.g. ImageNet [27], and recent methods based on deep convolutional neural networks (CNNs) [17, 13]. Following the trend in the well-studied object recognition and scene parsing tasks [25, 39], there is a growing interest in understanding the social, affective, and subjective aspects from the images [22], [29], [9]. Along this line, we question the feasibility of automatic ambiance inference from social media images of venues. Due to the inherent subjectivity of understanding ambiance, our study has an inseparable human interpretation part.

Ambiance is an important factor that affects the socializing experience in a venue [19, 16]. Therefore, spotting the visual cues that contribute to the ambiance of a place would enable a better understanding of the ambiance construct.

Understanding ambiance from images is challenging due to its subjectivity and is currently an issue of active research [29, 4]. Colors, textures, spatial layout and prior interpersonal experience and knowledge contribute to the ambiance concept. [3] showed that young people tend to prefer bright colors while adults prefer dark-colored environments. Similarly, age is an important factor in social environments [5]: while young people are likely to visit trendy places, older people are more often seen in conservative venues.

In this paper, we study ambiance of popular Foursquare (4SQ) places based on the image corpus used in [29]. This dataset consists of nearly 50K images and crowd-annotated place ambiance scores across 13 ambiance categories. Here, we focus only on *artsy* and *formal* categories.

Visual cues are reported to be used to infer ambiance in environmental psychology [11] e.g. the lens model by Brunswik. Thus, we hypothesize that the presence or type of certain objects may provide insights on ambiance, e.g. paintings on the wall for an artsy appearance. In this paper, we address the following two research questions:

1. What visual cues do crowdworkers associate with specific ambiances in social media images?
2. Are the outputs of deep networks trained on ambiance categories interpretable by crowdworkers? If so, do the learned visual patterns match with human-defined cues?

In the rest of the paper, first, we present the related work on ambiance in social computing and environmental psychology. Secondly, we describe the datasets used in this paper. Thirdly, we present our three-step methodology. Then

we give details on the experimental settings and results. Finally, we conclude by providing the insights from both our classification and crowdsourcing study.

2. Related Work

Social media platforms have enabled the collection of vast amounts of multimedia data about everyday life. Furthermore, online platforms to collect annotations of multimedia data via crowdsourcing have brought human perception into the picture. The combination of these two factors have opened the door for automatic analysis of subjective aspects of everyday life.

Ambiance in social computing. Thanks to available geo-tagged image datasets (e.g. [28]), various urban perception studies [20, 22, 24] analyzed outdoor places and cityscapes in terms of different subjective qualities, e.g. safety, wealth, uniqueness, beauty, etc. Similar studies on identifying the identity of a city [8, 38] focused on spotting discriminative elements of the scenes. In the perception of an observer, relating discriminative visual elements to ambiance follows Brunswiks lens model on environments [11].

To be able to identify artsy ambiance in venues, one can consider whether a venue looks interesting [7], has a certain style [15], or is memorable [14, 10] in an aesthetic sense. As shown in [14], when the memorability of a whole image is studied, a strong correlation is found among aesthetics and interestingness and human observers considered that beautiful and interesting images would be more memorable. In another study [10] that analyzes object-level memorability in images, predicted object category and saliency (in case of few objects in a scene) are stated to play an important role. Furthermore, people were found more memorable compared to other objects, e.g. vehicles and furniture.

The human factor in a scene is closely related to determining the ambiance of a place. As shown in the work of [12], human observers were able to guess the ambiance of a Foursquare venue based on the profile picture of visitors. Similarly, [26] showed that human observers may based their choices on facial visual cues indicating demographics. This motivates us to also investigate the human-related attributes in the scene in our crowdsourcing study on visual ambiance cues.

Moreover, physical layout, design and decorations are considered to be some of the main contributors of ambiance. Color, lighting, and style were shown to have strong influence on ambiance impressions in hotel lobbies [6], customer emotions [2], food choice [33], and post-dining intentions [19]. Thus, we included questions in our crowdsourcing task about lighting, color and organization of places.

Thanks to the crowdsourcing study in [29], an image corpus of popular Foursquare venues was annotated for 13 ambiance dimensions. This dataset enabled to study automatic ambiance inference based on traditional image features such as color and texture as well as deep representations from pretrained CNNs [30, 4]. Compared to the global prediction task from the whole image in [30], a two-step analysis is conducted in [4]: (1) semantic segmentation of objects in the scene, and (2) correlation analysis of the pixel percentage of the objects with the ambiance scores. However, in both studies, the utilized CNN models were pretrained on large benchmark datasets and were not adapted for the task. Furthermore, both previous studies lack a verification step at the end of the prediction task due to the unavailability of the groundtruth object or scene labels. In this paper, we work on the same dataset with two substantial differences, i.e. training a deep CNN model for ambiance prediction as well as localization of discriminative scene parts with the same model, and conducting a detailed crowdsourcing analysis on the visual cues.

Interpretation of deep networks. To understand the representations learned by deep CNNs, Zeiler et al. discuss how to visualize them via deconvolutional layers [35]. They also present a method called occlusion maps such that a sliding window in an image is occluded and the predicted label of the image by the CNN is checked to see whether that region is important to identify the correct label.

Simonyan et al. presented a simple gradient backpropagation approach for identifying the salient points of the objects with a single forward pass [32]. Compared to occlusion maps, it is computationally more efficient. However, this approach does not point out to the full object extent in general. Therefore, the authors use the output salient points from this approach as input to a classical background/foreground segmentation method for object segmentation in natural images.

Zhou et al. introduced class activation maps (CAM) [36] for capturing the extensions of objects and not only few salient points of objects. The CAM approach requires to introduce an average pooling layer to model structure. To avoid that, Grad-CAM [31] was proposed as a generalization of CAM. As such, it does not require modifying the CNN model to visualize the activation maps, and it can be applied to any type of neural networks, even to pretrained ones without the need of re-training. Following the related work on understanding CNN representations for social media and urban perception data [1, 23, 21], in this paper, we adopt the Grad-CAM approach for illustrating the discriminative parts of the scenes for a specific ambiance.

3. Datasets

In this paper, we use the dataset used in [29]. The dataset consists of venue images collected from Foursquare and crowdsourced ambiance judgments for a manually-chosen subset of images.These two data sources are explained in detail below. In the rest of this paper, we will use the terms place and venue interchangeably. Also note that, while the

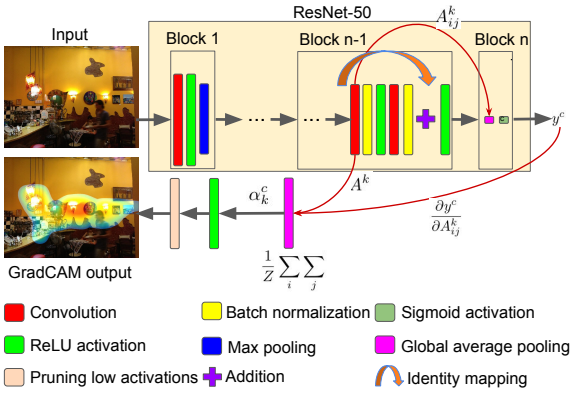


Figure 1: Illustration of the GradCAM method.

data was collected from Foursquare, the image content is similar to what is shared on many other social media sites (Facebook, Instagram, etc.) regarding views of places, patrons, and activities.

Foursquare 50K Dataset. This dataset consists of Foursquare (4SQ) images collected from 278 venues, for a total of 45,848 images with each venue having an average of 164 images [29]. The venues are mostly restaurants, cafes, bars, and clubs in six large-scale cities, i.e. Barcelona, Mexico City, New York City, Paris, Seattle, and Singapore. In this paper, we only considered a subset of this 50K corpus by following the same filtering procedure as in [4]. We will refer to this subset as *7K-resto* corpus in the rest of the paper. Specifically, this subset contains 7605 images that were predicted as one of the following labels: restaurant, stage, library, barbershop, cinema, grocery store, shoe shop, tobacco shop, bakery, and dining table. These are the top 10 most commonly predicted labels for the manually-chosen physical environment corpus by a GoogleNet [34] pretrained on the MIT-Places dataset [37]. The details of the manually-chosen physical environment data is given below.

Foursquare Physical Environment Dataset. This dataset consists of a subset of 3 images per venue. These images are manually-chosen such that the environment and the inner space of the venue can be observed from different angles. In the initial study of [29], these manually chosen images were found more informative of ambiance. By exposing the annotators to these manually chosen images, perception scores per venue were collected via online crowdsourcing using Amazon Mechanical Turk (MTurk). Each venue was annotated along 13 ambiance dimensions that are appropriately selected for indoor places. 10 annotations per place were collected in the range 1 (low) to 5 (high). Since these places were popular on 4SQ, overall the scores were high for positively phrased ambiances (e.g., *trendy*, *artsy*). In the rest of the paper we refer to the 4SQ manual physical environment dataset as *PhysEnv* dataset.

4. Methodology

To understand the ambiance of the venues according to the available image data, we followed three steps: 1) ambiance classification with deep residual convolutional neural networks, 2) visualization of discriminative visual cues via GradCAM, 3) analysis of the crowd perception on the generic ambiance visual cues as well as on the discriminative power and interpretability of GradCAM visualizations.

Ambiance Classification with Deep CNNs. To learn discriminative visual cues for the ambiance of a venue, we chose to transfer knowledge from an existing pretrained network through fine-tuning. We considered the state-of-the-art residual deep CNNs, specifically ResNet-50 model, that is pretrained on ImageNet data. In this step, we formalized the problem simply as binary classification. Thus, we replaced the final softmax activation in the pretrained model with a sigmoid activation and updated the weights of all the layers during fine-tuning. Our approach is equivalent to training a deep CNN from scratch with a *warm start*, i.e. starting from relevant and converged pretrained weights.

Visual Cue Visualization via GradCAM. To understand CNN representations, we used Grad-CAM [31]. With these methods, we visualized the salient and the discriminative parts of the ambiance labels.

Grad-CAM. As a generalization of Class Activation Mapping (CAM) [36], Gradient-weighted CAM (Grad-CAM) does not need a change in the CNN architecture [31].

Fig. 1 illustrates the GradCAM method. Essentially, first, the backpropagated gradients $\frac{\partial y^c}{\partial A^k_{ij}}$ are global-average-pooled and the importance weights α^c_k of feature map activations A^k are obtained:

$$\alpha^c_k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k_{ij}}, \tag{1}$$

These weights correspond to the re-trained weights w^c_k in CAM approach. Secondly, the linear combination of these weights and the feature map activations A^k are computed. Finally, this linear combination is passed through a ReLU activation so that only positive activation-gradient combinations are considered.

$$L^c = ReLU\left(\sum_k \alpha^c_k A^k\right). \tag{2}$$

In our case, to pay attention only to the most characteristic scene parts (visual cues), as a final operation, we eliminated the weak activation-gradient combinations (lower than 0.25) in the localization map.

Crowdsourcing Task. To assess the interpretability of the trained CNNs *qualitatively*, we performed a crowdsourcing study. Specifically, this study is a perceptual anal-

Figure 2: First part of the task design.

Figure 3: Second part of the task design.

ysis of the discriminative visual cues that the residual network (ResNet-50) captured for each ambiance class.

As a crowdsourcing platform, we used *Crowdflower*. This platform has *test questions* and a related *quiz mode* options to ensure the quality of the annotations by eliminating the annotators who provide unexpected answers (out of a predefined acceptable set of answers) to certain parts of the task. However, since our task is about assessing subjective qualities, we did not applied *quiz mode* or *test questions* in our task. To ensure the reliability of the results, we allowed only those crowdworkers who achieved *level-3* (highest level) qualification from their previous tasks on Crowdflower platform. Moreover, we set a minimum time to complete a task as one minute. We also set the maximum number of annotations per crowdworker as 20. Furthermore, to be consistent with the ambiance score annotation study of [29], we wanted to have similar demographics, thus we allowed only crowdworkers from the USA to perform our task. We set the payment as 10 cents (USD) for a task. 5 annotations were collected per task. Overall, we collected 500 annotations for 100 images and the total cost of our crowdsourcing study was 60 USD with the additional Crowdflower transaction fee (20%).

As opposed to crowdsourcing studies with every-day objects in natural images, we have the challenge of abstract and subjective concept of ambiance categories. Thus, we first launched a pilot task on the manually-picked images of the 10 places with highest avg. ambiance score from [29]. For each place, we picked one image and prepared *mock-up* heatmaps that highlighted the relevant visual cues in our consideration, i.e. paintings for artsy ambiance and tablecloth for formal ambiance. Moreover, we prepared detailed instructions and examples. Our focus was on whether non-experts' perception is aligned with the automatic discriminative models. From the pilot task, we observed that there was enough inter-rater agreement among the crowdworkers. Thus we decided to launch the main task for artsy and formal ambiances (50 places each).

Table 1: Test accuracies (%) obtained by training the ResNet-50 models on different corpora.

	artsy		formal
	PhysEnv	7K-resto	7K-resto
test accuracy	62.0 %	83.1 %	83.5 %

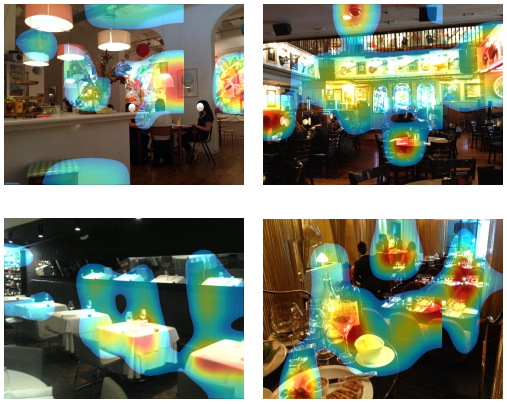


Figure 4: GradCAM visualizations as heatmaps (red to blue: high to low probability) for two artsy (top row), and two formal places (bottom row).

Our crowdsourcing task is composed of two main parts.

Part 1: Ambiance Visual Cue Identification. As shown in Fig.2, in the first part, the annotator observes an individual image from a place and answers ambiance-related questions such as type of the place (restaurant, cafe, bar, etc.), general feeling of the place (dark/bright, cluttered/organized, etc.), attire and age of the people in the image, and visual cues (presence of objects or type of objects) that contribute to the ambiance. The list of visual cues that might have an impact on the ambiance was gathered by the authors of this paper by checking the images of the 10 venues with highest average ambiance score from PyhsEnv corpus.

Part 2: Ambiance Discriminative Scene Part Assessment. Fig. 3 illustrates the second part of our crowdsourcing task. In this part, the annotator observes the visualization of the GradCAM method for the same image and rates whether it is a good visualization and main cues were captured in a scale of 7 ranging from "very poor" to "very good". We also ask the annotator to rate the missing or unexpectedly-highlighted visual cues in the heatmaps. Moreover, we ask the annotator to report as free text what was captured, missed or unexpectedly-marked in the heatmaps.

5. Experimental Settings and Results

Ambiance Classification. Based on the average ambiance scores obtained in the previous study [29], we labeled the images from the venues with average score above 3 (in the range of 1 to 5) as positive, and the ones with the average score below 3 as negative. Note that, among 278 places, only 49 places had above 3 average score for *artsy* ambiance whereas there were 227 *non-artsy* places. Similarly, there were 35 *formal* vs. 237 *non-formal* places. As the next step, we divided the places in the PhysEnv corpus into training, validation and test sets (around 80%-10%-10%). Based on this division of PhysEnv corpus, we also divided the 7K-resto subset of the 50K corpus by picking the images from these places. Next, we cropped the maximal square crops from each image and propagated the ambiance labels. To overcome the data imbalance, we oversampled the positive crops with a random combination of the standard image transformations, i.e. rotation, translation, and scaling. We applied nearest color filling if necessary. For the PyhsEnv corpus, we oversampled 4000 training, 1000 validation, and 1000 test image crops per class. Similarly, for the 7K-resto corpus, we oversampled 16000-2000-2000 crops per class respectively.

For both the data cases, we fine-tuned all the layers of the pretained ResNet-50 (learning rate 10^{-6}) for at least 100 epochs. As shown in Table 1, the models trained on the PhysEnv data for artsy ambiance reached 62.0% test accuracy. For the 7K-resto corpus, we obtained 83.1% and 83.5% for *artsy and formal* ambiances respectively.

Note that while the 7K-resto images contains visual cues that are not about the physical environment (selfies and food/drinks photos), this content is still informative of the ambiance of places (e.g. types of plates, cutlery, and glasses), and so the model captures more variety of the venue images and therefore outperform the models trained on the smaller PhysEnv corpus. Therefore, we decided to move forward with the models trained on 7K-resto corpus to visualize GradCAM responses in the next step.

Visualizations via GradCAM. As illustrated in Fig. 1, an input image is forwarded through the trained ResNet-50 model in a single pass, and the filter activations on the first convolution of last residual block were used to compute the heatmaps. In our experiments, we passed maximal square crops of the input image and then merged their heatmaps (by taking the maximum value in the intersection areas of two crops). Since GradCAM visualizes only the positive activations of an input, we picked 50 images (per ambiance) that had at least one square crop predicted as positive class.

Figure 4 shows some example visualizations. In artsy examples, the model bases its prediction on paintings or lighting, whereas in formal examples, tablecloth, wine glasses, people in formal attire (suits, etc.) were highlighted.

Analysis of Crowdsourced Ambiance Perceptions. We present the analysis of our crowdsourcing study that had two main parts below. Overall, 42 crowdworkers contributed to these results. Half of these workers completed

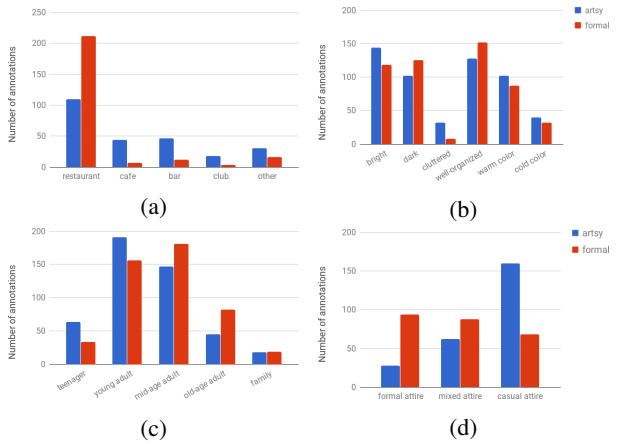


Figure 5: Part 1: Histograms of the annotations for venue type perception and general atmospheric feeling (top row), and age and attire of people in the scene according to the annotators (bottom row). Blue: artsy, red: formal.

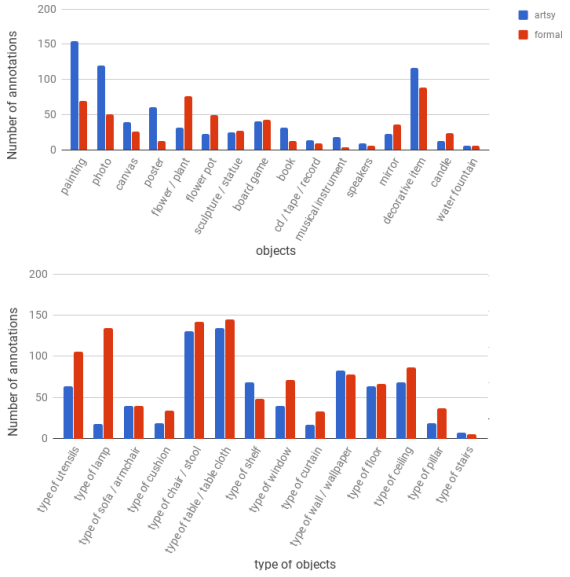


Figure 6: Part 1: Number of annotations for the objects (top plot) or the object types (bottom plot) that have an impact on ambiance.

the preset maximum limit of 20 tasks. The median number of annotations per worker was 15.

Part 1: Ambiance Visual Cue Identification. In the first part of our crowdsourcing study, the annotators marked the perceived type of the venue mostly as *restaurant* (more than 200 annotations out of 250 total annotations over 50 images) for the *formal* ambiance case as shown as red bars in the top-left plot of Fig. 5a. Similarly, in the same plot, over 100 annotations in 250 total annotations were *restaurant* as shown with the blue bars for the perceived venue type

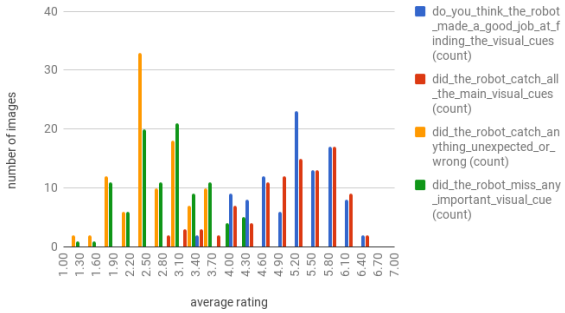


Figure 7: Part 2: Histogram of average ratings (scale from 1 to 7) on how crowdworkers perceived the GradCAM visualizations of ambiance labels.

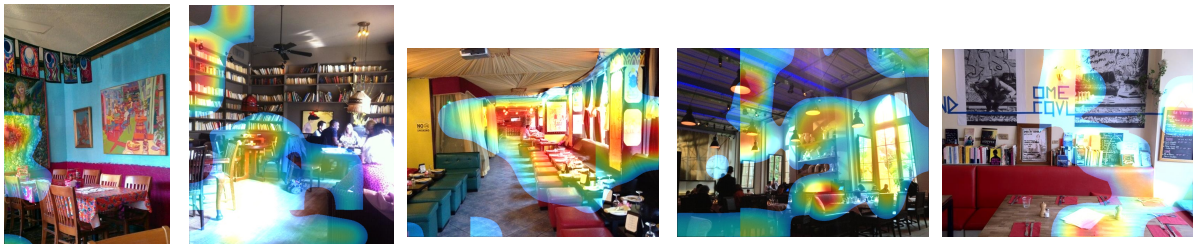
for the *artsy* ambiance. However, other type of venues, i.e. cafe, bar, club, and others were marked to be more likely to be *artsy* than *formal*. Moreover, after aggregating the venue type annotations per place (via winner-takes-it-all principle), we observed a clear dominance of *restaurant* type (49 out of 50) in the *formal* places. Although the *restaurant* type was found as dominant in the aggregated annotations of the artsy places as well (23 out of 50 places), other venue types (11 bar, 8 cafe, 4 club, and 4 other) were also noted in these selected 50 places.

Furthermore, Fig. 5b presents the number of annotations on the general feeling of a venue across illumination, color, and organization aspects. Following the trend in Fig. 5b, according to the aggregated markings, artsy places were found to be brighter (22 vs. 10) and more likely to have warm colors (8 vs. 5) than the formal places. Additionally, formal places were found more well-organized (20 vs. 11) and darker (15 vs. 9) than the artsy places in this subset of the PyhsEnv corpus.

About the customer age range and attire, Fig. 5c and 5d show the distribution of annotations for two ambiance cases. As reflected in the aggregated markings, young adults are more likely to be found in artsy places rather than middle-aged (34 vs. 12 in 50 images respectively) in this subset of the PyhsEnv corpus. On the other hand, the crowd agreed that middle-aged adults are more likely to be observed than young adults (27 vs. 20 in 50 places) in formal places.

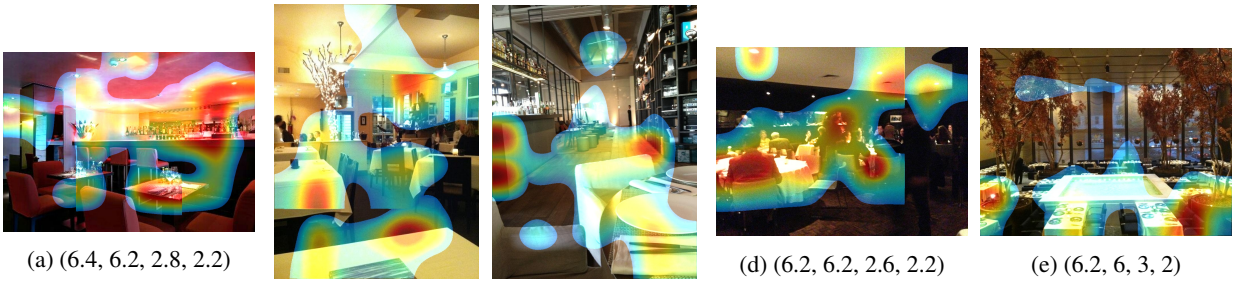
Fig. 6 shows the crowd annotations about the visual cues in the first part of our crowdsourcing task. Specifically, the top plot shows the number of annotations in case the *presence of an object* was perceived to have an impact on the venue ambiance by the crowd. Similarly, the bottom plot shows the number of annotations if the *type of the object* was found to influence the ambiance. We observe that the presence of a painting, a photo or a decorative item were important visual cues for artsy ambiance, whereas presence of a flower/plant, type of utensils and type of lamp were found as strong visual cues for formal ambiance. Furthermore, in-

(a) (6.4, 6.4, 1.6, 1.6) (b) (6.2, 6.2, 1.8, 1.8) (c) (6.2, 6, 2.6, 2.2) (d) (6.2, 6, 3.2, 2.6) (e) (6, 6, 3, 2.8)



(f) (3, 4.6, 4.2, 2) (g) (3.2, 4.6, 3.6, 3.4) (h) (3.2, 4.8, 4.2, 3.6) (i) (3.6, 3.6, 3.2, 3.4) (j) (3.6, 4.2, 3.6, 3)

Figure 8: Part 2: Ranked GradCAM visualizations by crowdworkers for artsy ambiance (top row: highest-rated, bottom row: lowest-rated from left to right). Average scores for four ratings were shown in parenthesis, i.e. "did the robot catch all the main cues?", "did the robot do a good job?", "did the robot miss any visual cues?", "did the robot catch anything unexpected or wrong?", respectively.



(a) (6.4, 6.2, 2.8, 2.2) (b) (6.2, 6.4, 2.8, 2.8) (c) (6.2, 6.2, 1.2, 1.2) (d) (6.2, 6.2, 2.6, 2.2) (e) (6.2, 6, 3, 2)

(f) (3, 4, 3.6, 3.2) (g) (3.2, 4.6, 3.4, 2.6) (h) (3.6, 3.6, 3.2, 3.2) (i) (4, 4.2, 3.6, 3.2) (j) (4.2, 4, 3.2, 3.2)

Figure 9: Part 2: Ranked GradCAM visualizations by crowdworkers for formal ambiance (top row: highest-rated, bottom row: lowest-rated from left to right). Average scores for four ratings were shown in parenthesis, i.e. "did the robot catch all the main cues?", "did the robot do a good job?", "did the robot miss any visual cues?", "did the robot catch anything unexpected or wrong?", respectively.

terior design and layout-related cues such as type of seats (chair/stool), type table/table cloth, type of wall/wallpaper, type of ceiling, type of floor, type of window were perceived as important cues for both ambiance cases.

Part 2: Ambiance Discriminative Scene Part Assessment. Fig. 7 illustrates the average ratings on four rating questions (scale: 1 to 7, where 1 corresponds to the most negative and 7 corresponds to the most positive response

Table 2: Part 2: The descriptive statistics of the ratings about GradCAM heatmap assessment.

	Did robot catch all main visual cues?		**Do you think robot did a good job?**		**Did robot miss any important visual cues?**		**Did robot catch anything unexpected or wrong?**	
	artsy	**formal**	**artsy**	**formal**	**artsy**	**formal**	**artsy**	**formal**
mean	5.024	5.22	5.236	5.236	2.752	2.692	2.544	2.432
std. dev.	1.637	1.519	1.474	1.490	1.576	1.629	1.528	1.506
first quartile	4	4	4	4	2	1	1	1
third quartile	6	6.75	6	7	4	4	4	4

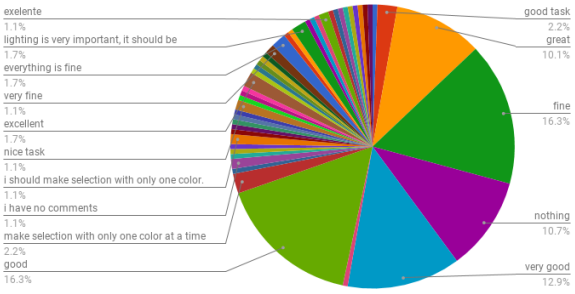


Figure 10: Standardized free-text feedback from annotators about our task design.

to the posed question.) in the second part of our crowdsourcing study. This plot shows that the crowd agreed on the accuracy and completeness of our GradCAM visualizations, since the average rating of *"Do you think the robot did a good job at finding the visual cues?"* and *"Did the robot catch all the main visual cues?"* questions (blue and red bars, respectively) were rated above 4 for most of the images, and had a median of 5 (in the range of [1,7]). Moreover, for only few images, our GradCAM visualizations were found to miss important visual cues or marking unexpected objects (most yellow and green bars are lower than 4 with a median of 2). The descriptive statistics of the ratings on the GradCAM visualization assessment are presented in Table 2. Furthermore, the GradCAM heatmaps that were ranked with the highest and lowest average scores on the positive questions (blue and red bars in Fig. 7) are illustrated in the top and bottom rows of Fig. 8 and 9 for artsy and formal ambiance cases, respectively. For instance, in the highest-rated artsy visualization in Fig. 8a, we can see that all the decorative items on the shelves, colorful lamps, and the large window was highlighted by GradCAM. However, in the lowest-rated visualization in Fig. 8f, the GradCAM method highlighted only one of the paintings partially and was insufficient to cover all the colorful paintings and the tablecloth. This resulted in a relatively low appreciation of the machine's work by the crowd (3.4 avg. rating out of 7 for "caught all main cues" and 4.2 rating for "missed some cues" questions). Similarly, in the highest-rated formal visualizations (top row of Fig. 9), all the tables/tablecloth, cutlery, wine glasses, and fancy lighting, i.e. spotlights on the ceiling or chandelier, are captured by GradCAM. However, for instance in Fig. 9g, GradCAM result is perceived as "not caught all the main cues" by the crowd (3.2 avg. score is on the negative side of the scale) due to not capturing all the table, cutlery, utensils, and chandelier. Furthermore, we observed that the visual cue markings in the first part are in agreement with the free text that the crowdworkers stated about the highlighted main cues by GradCAM method in the second part of our crowdsourcing task.

Fig. 10 illustrates the feedback from the annotators about our task design after few standardization of free-text, i.e. correcting typos and merging similar comments. Most of the annotators found our task as a "good task" whereas few people suggested to do the annotations with one-color version of the GradCAM heatmaps.

6. Conclusion

This paper investigated the automatic inference of ambiance of social media venues as well as the interpretability of the trained deep residual convolutional models according to human perception on the discriminative visual cue visualizations via GradCAM method. Based on our crowdsourcing study on these visualizations, for most of the cases, the learned deep models were able to capture the main visual cues contributing to the ambiance. One of the limitations of our work is the relatively small amount of labeled data we had for the positive examples for the ambiance categories we studied. With the availability of more physical environment relevant images and some filtering of food/drink and closeup selfie shots, the prediction and visualization results would be likely to improve. Overall, our automatic ambiance visual cue recognition results and crowdsourcing methodology to interprete and validate them are promising, and may motivate further studies on the understanding of additional ambiance categories and other subjective attributes of places in social media.

7. Acknowledgments

This work was supported by the Hasler Foundation (DCrowdLens project) and the Swiss National Science Foundation (Dusk2Dawn project).

References

[1] X. Alameda-Pineda, A. Pilzer, D. Xu, N. Sebe, and E. Ricci. Viraliency: Pooling local virality. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] R. Bell, H. L. Meiselman, B. J. Pierson, and W. G. Reeve. Effects of adding an italian theme to a restaurant on the perceived ethnicity, acceptability, and selection of foods. *Appetite*, 22(1):11–24, 1994.

[3] J. A. Bellizzi, A. E. Crowley, and R. W. Hasty. The effects of color in store design. *Journal of retailing*, 1983.

[4] Y. Benkhedda, D. Santani, and D. Gatica-Perez. Venues in social media: Examining ambiance perception through scene semantics. In *Proceedings of the 25th ACM International Conference on Multimedia, ACM, 2017*, 2017.

[5] B. Bishop and R. G. Cushing. *The big sort: Why the clustering of like-minded America is tearing us apart.* Houghton Mifflin Harcourt., 2009.

[6] 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*, 2006.

[7] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011.

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Trans. Graph.*, 2012.

[9] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*, pages 196–212. Springer, 2016.

[10] R. Dubey, J. Peterson, A. Khosla, M. Yang, and B. Ghanem. What makes an object memorable? In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1089–1097, 2015.

[11] R. Gifford, L. Steg, and J. P. Reser. Environmental psychology. In *IAAP Handbook of Applied Psychology*, pages 440–470. Wiley-Blackwell, 2011.

[12] 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.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1469–1482, 2014.

[15] Y. Jae Lee, A. A. Efros, and M. Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1857–1864, 2013.

[16] P. Kotler. Atmospherics as a marketing tool. *Journal of retailing*, 49(4):48–64, 1973.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012.

[18] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[19] Y. Liu and S. S. Jang. The effects of dining atmospherics: an extended mehrabian–russell model. *International Journal of Hospitality Management*, 2009.

[20] N. Naik, J. Philipoom, R. Raskar, and C. A. Hidalgo. Streetscore - predicting the perceived safety of one million streetscapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*, pages 793–799, 2014.

[21] L. S. Nguyen, S. Ruiz-Correa, M. S. Mast, and D. Gatica-Perez. Check out this place: Inferring ambiance from airbnb photos. *IEEE Transactions on Multimedia*, 2017.

[22] V. Ordonez and T. L. Berg. Learning high-level judgments of urban perception. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, pages 494–510, 2014.

[23] L. Porzi, S. Rota Bulò, B. Lepri, and E. Ricci. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 139–148, New York, NY, USA, 2015. ACM.

[24] D. Quercia, N. K. O'Hare, and H. Cramer. Aesthetic capital: what makes london look beautiful, quiet, and happy? In *Proc. CSCW*. ACM, 2014.

[25] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[26] M. Redi, D. Quercia, L. T. Graham, and S. D. Gosling. Like partying? your face says it all. predicting the ambiance of places with profile pictures. In M. Cha, C. Mascolo, and C. Sandvig, editors, *ICWSM*, pages 347–356. AAAI Press, 2015.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[28] P. Salesses, K. Schechtner, and C. A. Hidalgo. The collaborative image of the city: mapping the inequality of urban perception. *PloS one*, 8(7):e68400, 2013.

[29] D. Santani and D. Gatica-Perez. Loud and trendy: Crowdsourcing impressions of social ambiance in popular indoor urban places. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 211–220. ACM, 2015.

[30] D. Santani, R. Hu, and D. Gatica-Perez. Innerview: Learning place ambiance from social media images. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 451–455. ACM, 2016.

[31] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[32] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[33] N. Stroebele and J. M. De Castro. Effect of ambience on food intake and food choice. *Nutrition*, 2004.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[35] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[36] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.

[37] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

[38] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *ECCV*. 2014.

[39] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.