Streetscore - Predicting the Perceived Safety of One Million Streetscapes

Nikhil Naik naik@mit.edu
Jade Philipoom jadep@mit.edu
Ramesh Raskar raskar@mit.edu
César Hidalgo hidalgo@mit.edu

MIT Media Lab

Abstract

Social science literature has shown a strong connection between the visual appearance of a city's neighborhoods and the behavior and health of its citizens. Yet, this research is limited by the lack of methods that can be used to quantify the appearance of streetscapes across cities or at high enough spatial resolutions. In this paper, we describe 'Streetscore', a scene understanding algorithm that predicts the perceived safety of a streetscape, using training data from an online survey with contributions from more than 7000 participants. We first study the predictive power of commonly used image features using support vector regression, finding that Geometric Texton and Color Histograms along with GIST are the best performers when it comes to predict the perceived safety of a streetscape. Using Streetscore, we create high resolution maps of perceived safety for 21 cities in the Northeast and Midwest of the United States at a resolution of 200 images/square mile, scoring ∼1 million images from Google Streetview. These datasets should be useful for urban planners, economists and social scientists looking to explain the social and economic consequences of urban perception.

1. Introduction

How does the appearance of a neighborhood impact the health and behavior of the individuals that inhabit them? During the last decades numerous research efforts have explored this question. This research has shown an association between neighborhood disorder and criminal behavior through the well-known 'broken windows theory' [26, 11], but also an association between neighborhood disorder and health outcomes, such as the spread of STDs [2], the incidence of obesity [6], and rates of female alcoholism [12].

Until recently, most data on the physical appearance of urban environments was based on low throughput surveys [15, 19]. More recently, however, online data collections methods where humans evaluate images, using experts [7] or crowdsourcing [22], have increased the availability of urban perception data, but not to the extent needed to create global maps. In fact, even the most ambitious crowdsourcing efforts [22] have a limited throughput, being able only to rank images from a handful of cities at a resolution of less than 10 images per square mile. These constraints limit the possibility of using survey based methods to create global maps of urban perception.

The good news about crowdsourced studies is that they provide an ideal training dataset for machine learning methods building on scene understanding literature in computer vision. A trained algorithm, in turn, can be used to create high resolution maps of urban perception at required spatial and geographical scales.

Here, we demonstrate that a predictor trained using generic image features and the scores of perceived safety from a crowdsourced study can accurately predict the safety scores of streetscapes not used in the training dataset. We evaluate the predictive power of different image features commonly used for scene understanding and show that this predictor can be used to extend existing datasets to images and cities for which no evaluative data is available.

2. Image Ranking using Trueskill

We use publicly available data from the crowdsourced study by Salesses et al [22] to train our algorithm. In this study, participants in an online game were shown two Google Streetview images randomly chosen from the cities of New York, Boston, Linz and Salzburg (fig. 1-(a)). Participants were asked to choose one of the two images in response to the question: 'Which place looks safer?'. $7,872$ unique participants from 91 countries ranked $4,109$ images using $208,738$ pairwise comparisons (or 'clicks').

We convert these preferences to a ranked score for each image using the Microsoft Trueskill algorithm [8]. Trueskill uses a Bayesian graphical model to rate players competing in online games. In our case, we consider each image to be the winner of a two-player contest when it is selected in response to a question over another image. Each player's skill is modeled as a $N(\mu, \sigma^2)$ random variable, which gets updated after every contest. The update equations for a two-

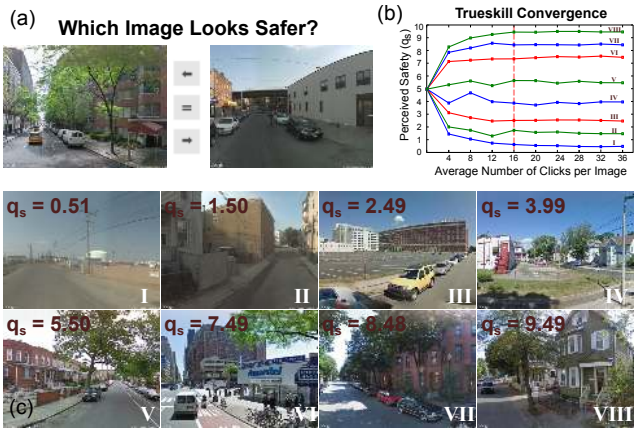

Figure 1. We convert the pairwise image comparisons obtained from a crowdsourced study (a) by Salesses et al. [22] to a ranked score using Trueskill [8]. (b) Trueskill converges to a stable score after ~16 clicks in our case. (c) The images are ranked on their perceived safety(q_s) between 0 and 10.

player contest [8] between players x and y, in which x wins against y, are as follows –

$$\begin{aligned}
\mu_x &\longleftarrow \mu_x + \frac{\sigma_x^2}{c} \cdot f\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right) \\
\mu_y &\longleftarrow \mu_y - \frac{\sigma_y^2}{c} \cdot f\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right) \\
\sigma_x^2 &\longleftarrow \sigma_x^2 \cdot \left[1 - \frac{\sigma_x^2}{c} \cdot g\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right)\right] \\
\sigma_y^2 &\longleftarrow \sigma_y^2 \cdot \left[1 - \frac{\sigma_y^2}{c} \cdot g\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right)\right] \\
c^2 &= 2\beta^2 + \sigma_x^2 + \sigma_y^2
\end{aligned} \tag{1}$$

where $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$ are trueskills of x and y. The pre-defined constant β represents a per-game variance, and ε is the empirically estimated probability that two players will tie. Functions $f(\theta) = \mathcal{N}(\theta)/\Phi(\theta)$ and $g(\theta) = f(\theta) \cdot (f(\theta) + \theta)$ are defined using the Normal probability density function $\mathcal{N}(\theta)$ and Normal cumulative density function $\Phi(\theta)$. Following [8], we use $(\mu = 25, \sigma = 25/3)$ as initial values for rankings for all images and choose $\beta = 25/3$ and $\varepsilon = 0.1333$.

In a two-player contest, Trueskill converges to a stable estimate of μ after 12 to 36 contests [8]. Figure 1-(b) shows that in our case, Trueskill estimates for μ are fairly stable after 16 clicks per image (we have on average, 35.92 clicks per image). Finally, we scale the scores to a range between 0 and 10 and denote these perceptual image scores, or 'Q-scores', mathematically as q_s. Going forward, we focus on the US and only use the 2920 images from New York and Boston.

3. Training a Computational Model

One would be inclined to believe that people's perception of safety is highly subjective. However, Salesses et al [22] demonstrate that the results obtained from their study are not driven by biases in age, gender or location of the participants, but by differences in the visual attributes of images. Hence creating a computational model for perceived safety based only on image features is feasible.

We draw upon previous work on image aesthetics whose goal is to predict the perceived beauty of photographs. Early research [4, 16] in this area was based on features extracted using rules from photography theory and psychology. Recently, however, task-independent generic image features have been found to be more effective for this purpose [18]. Similarly, we choose generic image features (e.g. [27, 14]) over rule-based features because it is not possible to create an exhaustive list of features for a set of images that is open and unknown.

Next, we describe our feature extraction process.

3.1. Image Feature Extraction

Figure 1-(c) shows eight images from our training dataset sorted from low to high scores. The typical high scoring image contains suburban houses with manicured lawns and streets lined with trees; while the typical low scoring image contains empty streets, fences, and industrial buildings. To predict their perceived safety, the image features need to capture this variance in appearance. Following Xiao et al. [27], we extract multiple generic image features which were shown to perform well for semantic scene classification, as summarized in Table 1. Specifically we extract GIST, Geometric Classification Map, Texton Histograms, Geometric Texton Histograms, Color Histograms, Geometric Color Histograms, HOG2x2, Dense SIFT, LBP, Sparse SIFT histograms, and SSIM.

Once we compute image features we train a predictor for the perceived safety of images (q_s) using Support Vector Regression as explained below.

3.2. Prediction using Support Vector Regression

To predict an image's perceived safety, we choose ν-Support Vector Regression (ν-SVR) [23]. Given input feature vectors $\mathbf{x}$ and their corresponding labels y, the goal of Support Vector Regression (SVR) with a linear kernel is to obtain a regression function $f(\mathbf{x})$ that approximates y, such that,

$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b, \qquad \mathbf{w}, \mathbf{x} \in \mathbb{R}^N, b \in \mathbb{R} \tag{2}$$

SVR tries to control both the training error and the model complexity, by minimizing the following function –

Feature	Computation Procedure
GIST [21]	Filterbank outputs (8 orientations at 5 different scales) are averaged on a 5×5 grid.
HOG2x2 [3]	124-dimensional 2x2 HOG descriptors are quantized into 300 visual words. Three-level spatial pyramid histograms are constructed and compared using histogram intersection.
Dense SIFT [13]	SIFT descriptors are extracted in a flat window at two scales on a regular grid at steps of 5 pixels. The descriptors, stacked together for each HSV color channel, are quantized into 300 visual words. Spatial pyramid histograms are used for kernels.
LBP [20]	Histograms of local binary patterns.
Sparse SIFT [25]	SIFT features at Hessian-affine interest points are clustered into dictionaries of 1,000 visual words. Two histograms of soft-assigned SIFT features are computed.
SSIM [24]	Correlation map of a 5×5 patch in a 40 pixel window is quantized into 3 radial and 10 angular bins. Spatial histograms constructed from 300 visual words of descriptors.
Texton Histograms	A 512-dimension histogram is built from a textonmap [17] of each image obtained from a universal texton dictionary.
Color Histograms	A joint $4 \times 4 \times 14$ histogram of color in *CIELab* color space for each image.
Geometric Classification Map	Histograms of geometric class probabilities [9] for ground, vertical, porous and sky.
Geometric Texton Histograms	Texton Histograms for each geometric class weighed by class probabilities.
Geometric Color Histograms	Color histograms for each geometric class weighed by class probabilities.

Table 1. Computation procedure for image features used to predict perceived safety.

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{K}\sum_{i=1}^{K}|y_i - f(\mathbf{x}_i)|_\epsilon$$
$$|y - f(\mathbf{x})|_\epsilon = \max\{0, |y - f(\mathbf{x})| - \epsilon\}, \qquad \epsilon > 0 \tag{3}$$

where the value of ϵ—chosen a priori—determines the desired accuracy of approximation.

The key idea of ν-SVR [23] is that by choosing ν instead of committing to a specific value of ϵ we guarantee that the number of predictions with an error more than ϵ is smaller than ν. This is achieved by solving the following constrained minimization problem –

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \left(\nu\epsilon + \frac{1}{K}\sum_{i=1}^{K}(\xi_i + \xi_i^*)\right) \\ \text{subject to} \quad & \Big((\mathbf{w}\cdot\mathbf{x}_i) + b\Big) - y_i \le \epsilon + \xi_i \\ & y_i - \Big((\mathbf{w}\cdot\mathbf{x}_i) + b\Big) \le \epsilon + \xi_i^* \\ & \epsilon \ge 0,\ \xi_i^* \ge 0 \end{aligned} \tag{4}$$

3.3. Performance Evaluation

We now evaluate the performance of these features by training an SVR for each feature. We use the Coefficient of Determination (R^2) between true scores q_s and predicted scores $\hat{q}_s$ to evaluate the accuracy of a regression model. R^2 is a quantitative measure for the proportion of total variance of true data explained by the prediction model. It is defined as

$$R^2 = 1 - \frac{\sum_i (q_s^i - \hat{q}_s^i)^2}{\sum_i (q_s^i - \bar{q_s})^2} \qquad \text{where} \quad \bar{q_s} = \tfrac{1}{|q_s|}\textstyle\sum_i q_s^i \tag{5}$$

We train the SVR using *libsvm* [1] and choose the linear kernel with following parameters: , $C = 0.01$, and $\nu = 0.5$. We determine the optimal C and ν using a grid search to minimize the R^2 of prediction over 5-fold cross-validation.

The performance of individual features for Q-score prediction is summarized in Fig. 2-(a). Geometric Texton Histograms perform the best ($R^2 = 0.4826$), followed by GIST ($R^2 = 0.4339$). It is important to note that HOG2x2 and Dense SIFT, which are the top performing features for generic scene classification [27], do not perform as well, with $R^2 = 0.3841$ and $R^2 = 0.4283$ respectively. A combination of all features gives $R^2 = 0.5676$.

Feature Selection : Our goal is to develop a trained predictor that can generate a very high resolution dataset of people's perception of urban environments. To reduce the computational cost of feature extraction from new images, we sought to choose the best three features using feature selection. We use 'forward selection' for this purpose. In this method, starting with an empty set, we iteratively add one feature at a time, which increases R^2 the most, over 5-fold cross-validation. This process gives us Geometric Texton Histograms, GIST and Geometric Color Histograms, in that order, as our three best features, with a combined $R^2 = 0.5365$ (Fig. 2-(b)). It is interesting to note that Geometric Color Histograms provide the most performance improvement after Geometric Texton Histograms and GIST, even though the individual performance of Geometric Color

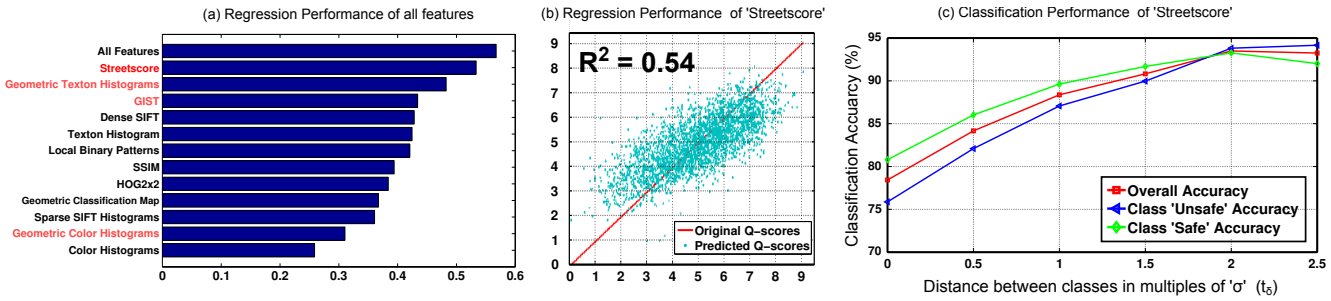

Figure 2. We analyze the performance of commonly used image features (a) for predicting perceived safety (q_s). Choosing the best performing features using forward selection, we train the 'Streetscore' predictor and analyze its performance for both regression (b) and binary classification (c).

Histograms is not impressive ($R^2 = 0.3103$). This shows that colors are an important dimension in predicting perceived safety as they provide information that is different from the one contained in textons and GIST.

We refer to a predictor trained using Geometric Texton Histograms, GIST and Geometric Color Histograms as the 'Streetscore' predictor and analyze its performance for binary classification as additional performance evaluation.

Binary Classification : For binary classification we use 'low' (t_l) and 'high' (t_h) thresholds to label images in our test set as 'unsafe' or 'safe' according to their Q-scores. We study classification accuracy as a function of $t_\delta = t_h - t_l$, where $t_l = \bar{q_s} - t_\delta/2$ and $t_h = \bar{q_s} + t_\delta/2$ and $\bar{q_s}$ is the average q_s of the test set (Fig. 2-(c)). For $t_\delta = 0$ (i.e. $t_l = t_h = \bar{q_s}$) the accuracy of our classifier is 78.42%, whereas for $t_\delta = 2 \cdot \sigma_{q_s}$ the accuracy is as high as 93.49%, over a $5-$fold cross-validation. We note that the accuracy of a random predictor would be 50%.

The robust performance of Streetscore in both classification and regression on this challenging dataset demonstrates that the computed features have good explanatory power for prediction of perceived safety. In the next section, we analyze the performance of Streetscore in different geographical regions of the United States.

4. Perception Maps

To use Streetscore for creating global maps of urban perception, we need to determine the generalization performance of the predictor in terms of geographical distance, that is, determine the spatial range for which we expect the predictions to hold. Are images from New York and Boston good enough to predict the perceived safety of images from all over the world? Or is the predictor trained using these images applicable only in a limited geographical region around New York and Boston? Intuitively, the external validity of the predictor would depend on the similarity between architectural styles and urban planning of the cities being evaluated. There are, however, no quantitative studies on measuring similarities along these axes. Therefore, we use the average median income of a city as a naïve metric to validate the accuracy of Streetscore for cities not in the original crowdsourced study.

First, we use StreetScore to create perception maps for 27 cities across the United States using images densely sampled from Google Streetview at 200 images/square mile (Fig. 3-(a)), scoring more than 1 million images. Then, using 9 cities which lie inside a 200 mile radius of New York City , we compute a linear fit (L_{qi}) between the mean Q-score ($\bar{q_s}$) of a city and its average median family income (I_c) according to the 2010 census (Fig. 3-(b)). We find a strong correlation between the two variables (Pearson Correlation Coefficient = 0.81).

This linear regression helps us determine some loose bounds on the geographical region over which the algorithm can be applied given the current training set. Figure 3-(c) shows a plot the residual from L_{qi} of cities as a function of their distance to New York City. This exercise shows that the mean absolute error (MAE) is large for cities in Arizona, California and Texas, indicating that the algorithm does not perform well in these regions. Based on this regression we propose a tentative bound of 1100 miles for the validity of our trained predictor. Certainly we do not expect this radius of validity to be the same in different parts of the world, where architectural gradients might be different. Nevertheless, even if these ranges vary by as much as a factor of two, they would indicate that an algorithm trained with a few images from a given area can be used to map a significant area of its surroundings, which is useful to know for future crowdsourced studies.

Figure 5 shows perception maps for six cities created using Streetscore. Our goal is to create such maps for every city in the Northeast and Midwest of United States and make them available through an interactive website. This dataset should help urban planners, economists and social scientists who are increasingly using data-driven techniques to study the social and economic consequences of urban environments.

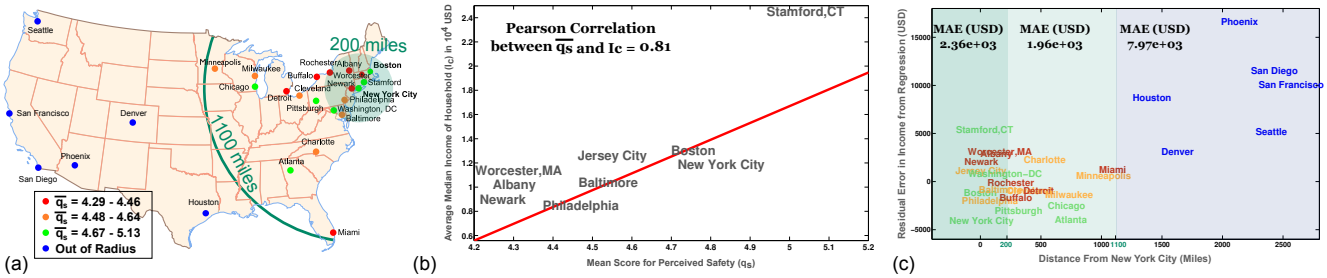

Figure 3. We evaluate the generalization ability of Streetscore in terms of geographical distance by scoring streetscapes from 27 cities across the United States (a). By using a linear fit (b) between the mean Q-score ($\bar{q}_s$) of a city and its average median family income (I_c), we calculate the residual for cities as we move farther away from New York (c) and establish a loose bound of 1100 miles from New York for our predictor.

5. Discussion and Limitations

Research in visual analysis beyond semantics, such as interestingness [5], beauty [4] and memorability [10], has been a recent topic of interest in computer vision. Our work helps expand this research by focusing on the evaluative dimensions of streetscapes, such as their perceived safety, liveliness and character. As a tool, Streetscore can help urban planners construct better cities. Yet, Streetscore also points to new research directions for the vision community. For instance, a fruitful area of research would involve identifying the objects and features that help explain the evaluative dimensions of a streetscape. Also, by classifying streetscapes evaluatively, it should be possible to advance data-driven rendering techniques that are tuned based on derived evaluative criteria (such as a place that is lively but harmonious).

Our results also show some important limitations. Since it is challenging to encode all the variance of the rich and complex visual world in a limited training dataset, the predictor fails when evaluating images with unusual visual elements, such as atypical architecture and stylistic elements like colorful graffiti (Fig. 4). Online learning methods which harness the information from user interaction to improve predictors can be used to overcome this limitation.

6. Conclusion

The visual appearance of urban environments can have strong effects on the lives of those who inhabit them. Hence it is important to understand the impact of architectural and urban-planning constructs on people's perception of streetscapes.

However, quantitative surveys of visual perception of streetscapes are challenging to conduct and as a result, existing surveys are limited to a sparse sampling of data from a few cities. In this paper, we demonstrate that a machine learning method trained using image features from a small dataset of images from New York and Boston labeled by a crowd, can be used to create 'perception maps' of 21 cities from United States at a resolution of 200 images/square mile.

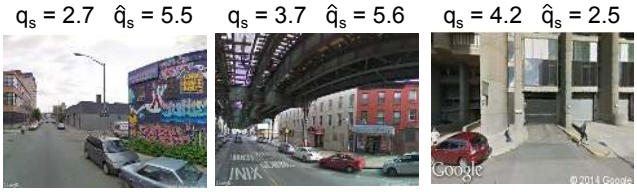

Figure 4. Failure cases: Streetscore can produce significant errors when evaluating images with rare visual elements not encountered in the training set, such as colorful graffiti, overpasses and modern architecture. (q_s – True Q-score, $\hat{q_s}$ – Predicted Q-score)

In conclusion, we present a novel computational tool for measuring perceived safety of cities which should inspire more research in quantitative analysis of cities and their impact on its inhabitants. In particular, further research on impact of different visual elements of streetscapes on perception can directly influence the urban planning and architecture community.

Acknowledgements

We would like to thank Daniel Smilkov for his help with the Trueskill algorithm and Michael Wu for his help in creating perception maps. We acknowledge support from the MIT Media Lab Consortia and the Google Living Labs Tides Foundation.

References

[1] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 3

[2] D. A. Cohen, K. Mason, A. Bedimo, R. Scribner, V. Basolo, and T. A. Farley. Neighborhood physical conditions and health. *American Journal of Public Health*, 93(3):467–471, 2003. 1

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Vision and Pattern Recognition*, pages 886–893, 2005. 3

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. 2006. 2, 5

[5] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Computer Vision and Pattern Recognition*, pages 1657–1664, 2011. 5

[6] A. Dulin-Keita, H. K. Thind, O. Affuso, and M. L. Baskin. The associations of perceived neighborhood disorder and physical activity with obesity among african american adolescents. *BMC public health*, 13(1):440, 2013. 1

[7] P. Griew, M. Hillsdon, C. Foster, E. Coombes, A. Jones, and P. Wilkinson. Developing and testing a street audit tool using google street view to measure environmental supportiveness for physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 10(1):103, 2013. 1

[8] R. Herbrich, T. Minka, and T. Graepel. Trueskill: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576, 2006. 1, 2

[9] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008. 3

[10] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Computer Vision and Pattern Recognition*, pages 145–152, 2011. 5

[11] K. Keizer, S. Lindenberg, and L. Steg. The spreading of disorder. *Science*, 322(5908):1681–1685, 2008. 1

[12] M. A. Kuipers, M. N. van Poppel, W. van den Brink, M. Wingen, and A. E. Kunst. The association between neighborhood disorder, social cohesion and hazardous alcohol use: A national multilevel study. *Drug and alcohol dependence*, 126(1):27–34, 2012. 1

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Vision and Pattern Recognition*, pages 2169–2178, 2006. 3

[14] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *IEEE Computer Vision and Pattern Recognition*, pages 2735–2742, 2012. 2

[15] K. Lynch. *The image of the city*, volume 11. the MIT Press, 1960. 1

[16] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, pages 83–92. ACM, 2010. 2

[17] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International journal of computer vision*, 43(1):7–27, 2001. 3

[18] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE International Conference on Computer Vision*, pages 1784–1791, 2011. 2

[19] J. L. Nasar. *The evaluative image of the city*. Sage Publications Thousand Oaks, CA, 1998. 1

[20] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 3

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 3

[22] 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. 1, 2

[23] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000. 2, 3

[24] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Computer Vision and Pattern Recognition*, pages 1–8, 2007. 3

[25] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470–1477, 2003. 3

[26] J. Q. Wilson and G. L. Kelling. Broken windows. *Atlantic monthly*, 249(3):29–38, 1982. 1

[27] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 2, 3

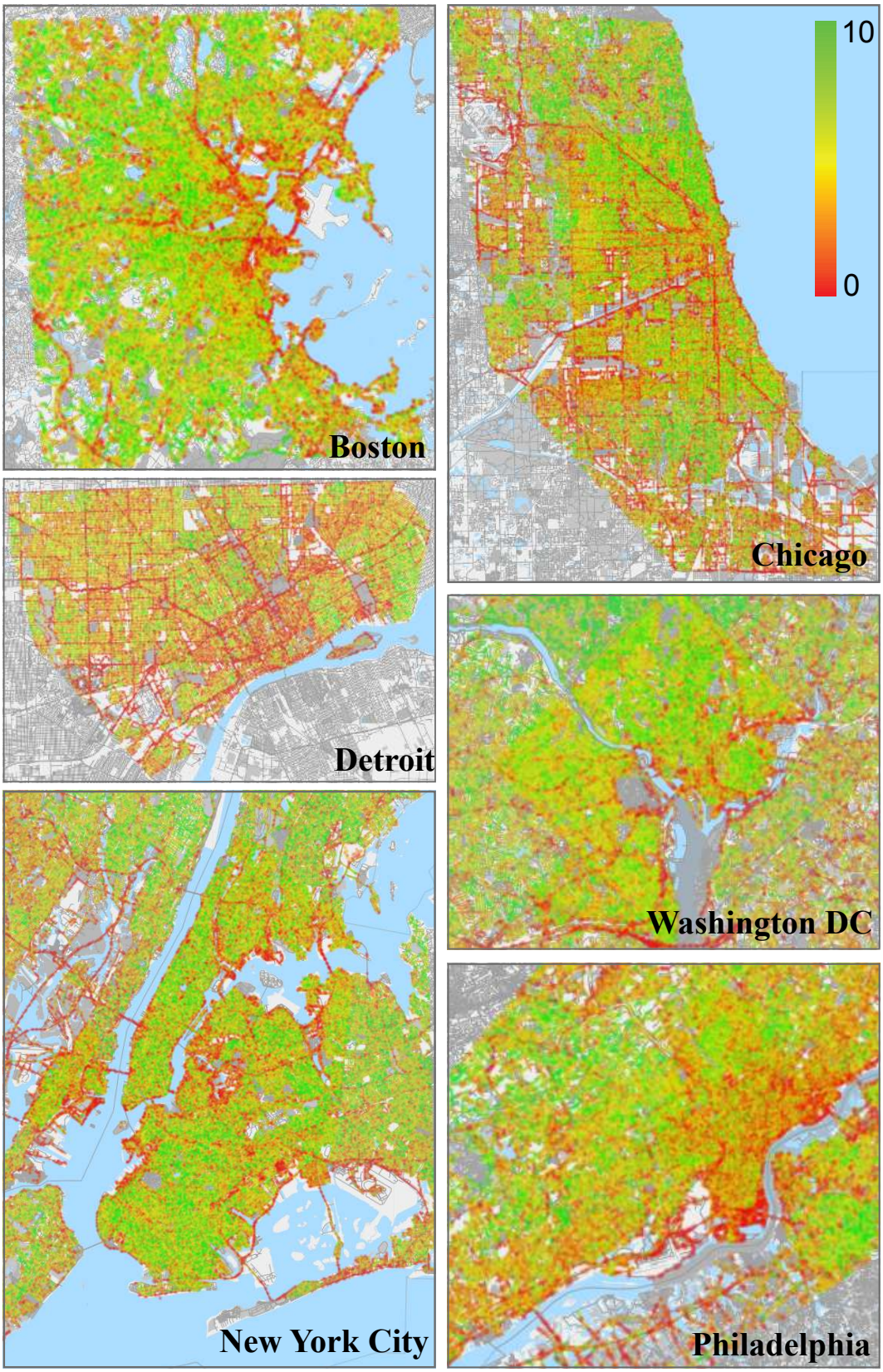

Figure 5. Perception maps for 6 cities at 200 images/square mile. An online interactive dataset of perception will be a useful tool for scientists studying social and economic impact of urban environments. (Note : the maps are at different scales)