Learning High-Level Judgments of Urban Perception

Vicente Ordonez and Tamara L. Berg

University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
{vicente,tlberg}@cs.unc.edu

Abstract. Human observers make a variety of perceptual inferences about pictures of places based on prior knowledge and experience. In this paper we apply computational vision techniques to the task of predicting the perceptual characteristics of places by leveraging recent work on visual features along with a geo-tagged dataset of images associated with crowd-sourced urban perception judgments for wealth, uniqueness, and safety. We perform extensive evaluations of our models, training and testing on images of the same city as well as training and testing on images of different cities to demonstrate generalizability. In addition, we collect a new densely sampled dataset of streetview images for 4 cities and explore joint models to collectively predict perceptual judgments at city scale. Finally, we show that our predictions correlate well with ground truth statistics of wealth and crime.

1 Introduction

Sense of place is a feeling or perception held by people about a location. It is often used to refer to those characteristics that make a place unique or foster a sense of belonging, but may also refer to characteristics that are not inherently positive such as fear [31].

In this paper we apply computer vision techniques to predict human perceptions of place. In particular we show that – perhaps surprisingly – it is possible to predict human judgments of safety, uniqueness, and wealth of locations with remarkable accuracy. We also find that predictors learned for one place are applicable to predicting perceptions of other unseen locations, indicating the generalizability of our models. Additionally, we explore models to jointly predict perceptions coherently across an entire city. Finally, we also find good correlations with ground truth statistics of crime and wealth when predicting on a more densely sampled set of images.

The world, or even a single city, is a large continuous evolving space that can not be experienced at once. The seminal work of Lynch, *The Image of the City* [19] was influential in urban design and the approach of social scientists to urban studies. Of course, collecting human judgments is a time consuming and costly process. With accurate computational prediction tools, we could extend human labeled data of a place to nearby locations or potentially the entire world, thus enabling social scientists to better understand and analyze public

D. Fleet et al. (Eds.): ECCV 2014, Part VI, LNCS 8694, pp. 494–510, 2014.

Fig. 1. Our goal is to learn the human perception of safety, wealth, and uniqueness for street level images. Human judgments agree that the image shown on the left is safer than the image shown on the right.

perceptions of places. Additionally, there are many potential applications of our method such as answering important questions that people might have about a place. For example, what areas should I avoid on my visit to NYC? In what neighborhoods in Chicago might I like to buy a house? Which blocks of Boston are the most unique?

Most computer vision algorithms related to places have focused on tasks like scene classification, (e.g. [16,26,33,29,18]) or parsing scene images into constituent objects and background elements (e.g. [30,10,15,32]). But, places are about much more than semantics. People perceive different qualities about a place, e.g. whether it is a safe place, an interesting place, or a beautiful place.

These notions are related to recent work on attributes, especially on predicting attributes of scenes [24]. Attributes such as scary, soothing, and stressful in the SUN Attribute dataset [24] are related to perceptual characteristics of safety, but are collected for a very different type of data. Our goal is somewhat different; while the SUN Attribute dataset consists of general internet images collected from Flickr, we look at streetview photos sampled densely across multiple cities (see Fig 2 for a comparison). In addition, past work on scene attribute recognition predicts attributes of images independently for each image. We take an approach that predicts attributes of all images within a location jointly using a graph based framework. Since images taken in nearby locations usually have similar perceptual characteristics this improves perceptual characteristic prediction performance. Finally, we also look at predicting perceptions at a much larger scale, e.g. on image sets spanning entire cities.

Our approach learns from a large data set collected by the *Place Pulse* project [28]. This dataset consists of 2920 streetview images of NYC and Boston. Ratings are collected from people regarding their perceptions of safety, uniqueness, and wealth. We train models for both classification (Sec 4.2, predicting e.g. which parts of a city are most or least safe), and regression (Sec 4.3, directly predicting perceptual ratings). Our quantitative evaluations demonstrate reliable performance for both tasks when training and testing on images from the same city. In addition, we experiment with training on images collected from one city and testing on images of another city and show good generalizability. Qualitative results also show that our learned models can predict which neighborhoods are most safe within a city. In addition to the original dataset, we

Fig. 2. Left: Sample images of *unsafe* street images. **Right:** Sample *scary* images from the SUN Attributes dataset [24]. Note, the distinct differences in types of image content between the collections.

collect additional photos for prediction (Sec 3.2) by densely sampling streetview images of NYC (8863 photos), and Boston (9596 photos), and 2 locations not in the original dataset – Chicago (12502 photos) and Baltimore (11772 photos). Finally, we show that our predictions of safety correlate well with statistics about crime and wealth in the 2 new cities, Chicago and Baltimore (Sec 6).

The main contributions of our paper are:

- Classification and regression models to predict human perceptions of the safety, uniqueness, and wealth depicted in images of places.
- Models to jointly predict perceptual characteristics of entire cities.
- Experiments demonstrating that perceptual characteristics of places can be predicted effectively when training and testing on the same city and when training and testing on different cities.
- Maps visualizing perceptual characteristics densely predicted over cities.
- Experimental evidence showing correlation between perceptual predictions and crime and wealth statistics.

2 Related Work

We discuss here several lines of research related to this work. We would also like to acknowledge a few concurrent efforts in perceptual prediction using urban data, most notably Naik *et.al.* [21], Arietta *et. al.* [2], Quercia *et.al.* [27] and Khosla *et. al* [14].

Scene Recognition & Reconstruction: There has been a lot of progress in scene recognition in recent years [16,26,33,24]. However, this research has mainly focused on scene categorization [16,26,33]. and recently on recognizing attributes of scenes [24]. Our task is somewhat different, trying to estimate human perceptions of place both of individual photos and in a coherent manner across larger extents, such as across an entire city. Additionally, rather than looking at all scene images, we focus on outdoor street level images of cities. For this task, there seem to be strong visual cues related to our high level knowledge

and experience with places. Content cues that may be related to perception of place include paintings on the walls (certain types of graffiti), presence or absence of green areas, presence of metallic fences and other objects, or amount and type of clutter. Another area of research related to place looks at reconstructing 3d models of scenes [1][9]. Recent methods operate at city scale. Our work could help put a semantic layer on top of these efforts by adding perceptual information to geometric models of places.

Geo-Locating Images. One previous related computer vision application is that of automatic image localization [11,34]. The work of Hays and Efros [11] uses a data-driven approach to predict image location (latitude and longitude) based on a collection of millions of geo-tagged images from Flickr. Later work from Zamir *et.al* [34] uses Google Street View images for this purpose. While these methods attempt to guess where a picture was taken, we try to predict aspects of the picture itself. In Hays and Efros [11] the authors also demonstrate that other meta-information such as population and elevation can be estimated based on geo-location predictions. Our work is similar in spirit in that we want to predict meta-information about images, but computes the prediction directly from the image content rather than using outside information such as elevation or population maps.

Perceptual Tasks: There has been recent interest in the vision community on predicting perceptual characteristics of images. Related tasks include predicting the aesthetic quality of images [6][20][12], discovering mid-level representations that are distinctive to a city [7] or to a style of object [17], and efforts to predict the memorability of images [12]. The most relevant to our work is the aesthetics task since it mimics the positive and negative nature of photos also present in predicting the safety of a location. For aesthetics, various approaches have been tried, including attribute based methods which train aesthetics classifiers based on the outputs of individual high level attribute detectors [6]. Though attribute based methods are intuitive, later work from Marchesotti *et.al* [20] found that generic image descriptors in combination with appropriate encoding methods can also produce state-of-the-art results. We use this insight to build our feature representations using recent state of the art image descriptors, in particular fisher vector (FV) encodings [25] and DeCAF convolution network based features [8].

3 Data

We use two main data sources in our work: a) the *Place Pulse 1.0* dataset collected by Salesses *et.al* [28] and labeled using crowdsourcing (Sec 3.1), and b) a larger street view dataset we collected for this work (Sec 3.2).

3.1 Place Pulse 1.0

We use the publicly available images from the *Place Pulse 1.0* dataset [28]. This dataset contains 1689 streetview images sampled across New York City and

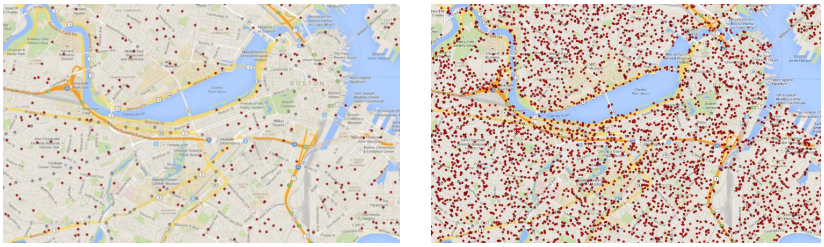


Fig. 3. The left image shows the sampling locations for the Place Pulse v1.0 dataset and the right image show the sampling locations of our unlabeled dataset for a zoomed-in section of the Boston/Cambridge area

1231 images of Boston. For each image in the dataset the authors provide meta-information related to location – geo-tags of latitude longitude – and camera rotation information. Each image i also comes with aggregated human judgment scores of perceived safety ($q_{i,s} \in Q_s$), uniqueness ($q_{i,u} \in Q_u$) and wealth/class ($q_{i,w} \in Q_w$). The locations in the dataset were randomly sampled across each city, with the exception of some locations for which there are multiple different views.

Perception scores for the 3 measures were collected via crowdsourcing using a website created for this purpose. On this website, a user is presented with two images side-by-side and asked to answer a relative perceptual judgment question, e.g. "Which place looks safer?". The user could select either the left or right image or tie. The goal of this project was to compute 3 scores for each image in the dataset $Q_s = \{q_{i,s}\}$, $Q_u = \{q_{i,u}\}$, $Q_w = \{q_{i,w}\}$ corresponding to safety, uniqueness, and wealth respectively. Due to practical considerations (limited numbers of users), not all possible pairs of images for a given city could be directly compared. Instead, the authors merged the pairwise rankings into an overall ranking by taking into account the relative judgments of the images against which each image was compared. This problem is a direct analog to the notion of "strength of schedule" [23] in sport matches.

Perceptual scores $q_{i,k}$ for perception type $k \in \{s, u, w\}$ for image i are:

$$q_{i,k} = \frac{10}{3}(W_{i,k} + \frac{1}{w_{i,k}} \sum_{j_1=1}^{w_{i,k}} W_{j_1,k} - \frac{1}{l_{i,k}} \sum_{j_2=1}^{l_{i,k}} L_{j_2,k} + 1) \tag{1}$$

$$W_{i,k} = \frac{w_{i,k}}{w_{i,k} + l_{i,k} + t_{i,k}} \quad , \quad L_{i,k} = \frac{l_{i,k}}{w_{i,k} + l_{i,k} + t_{i,k}} \tag{2}$$

Where the counts $w_{i,k}, l_{i,k}, t_{i,k}$ denote the number of times the image i won, lost, or tied compared to other images for perception metric k. The constant $(\frac{10}{3})$ was selected so that the output scores fall in the range $0 - 10$.

3.2 External Dataset

We additionally collect a much larger dataset of geo-tagged images for New York (8863 images) and Boston (9596 images), as well as for two new cities, Baltimore

(11772 images) and Chicago (12502 images). To collect this dataset, we use the Google Street View API to sample images from random locations within the boundaries of each city. To provide a better idea of the scale and coverage of our extended dataset, in Figure 3 we show side by side sampled locations for a zoomed-in area of Boston from the *Place Pulse 1.0* dataset (left) and our more densely sampled dataset (right). This denser sampling will allow us to generate urban perception maps and analysis at more detailed resolutions.

4 Predicting Urban Perceptions

We model and evaluate prediction of urban perceptions in two tasks, as a classification problem (Section 4.2), and as a regression problem (Section 4.3). First we describe the image representations used in these tasks (Section 4.1).

4.1 Image Representation

Since the seminal work of Oliva and Torralba on modeling the spatial envelope of the image [22], there have been several proposals for scene representations that leverage spatial information. The recent work of Juneja *et. al.* [13] presents a benchmark of several scene representations, including both low-level feature representations and mid-level representations. They find that using low-level features with rich encoding methods like Fisher vectors [25] can produce state-of-the-art results on challenging scene recognition problems.

For our work we evaluate three feature representations: Gist [22], SIFT + Fisher Vectors [25], and the most recent generic deep convolutional activation features (DeCAF) of Donahue *et. al.* [8]. For SIFT-FV we compute the SIFT features densely across five image resolutions, then perform spatial pooling by computing the FV representations on a 2x2 grid over the image and for the whole image. We build a visual dictionary with 128 components using Gaussian Mixture Models. Additionally, we use the rootSIFT variant and adopt other recommendations from Chatfield *et. al.* [4]. For the DeCAF features we use the output of the sixth convolutional layer in the neural network.

4.2 Classification

We set up the classification problem protocol in a similar manner to that used in image aesthetics tasks [20,5,6], where one tries to discriminate between images with high perceptual scores from images with low perceptual scores (commonly used in perceptual tasks since the scores of images middling perception values may not be stable across people). For classification, we define the binary labels $y_{i,k} \in \{1, -1\}$ for both training and testing as:

$$y_{i,k} = \begin{cases} 1 & \text{if } \; rank(q_{i,k}) \text{ in the top } \delta\% \\ -1 & \text{if } \; rank(q_{i,k}) \text{ in the bottom } \delta\% \end{cases} \tag{3}$$

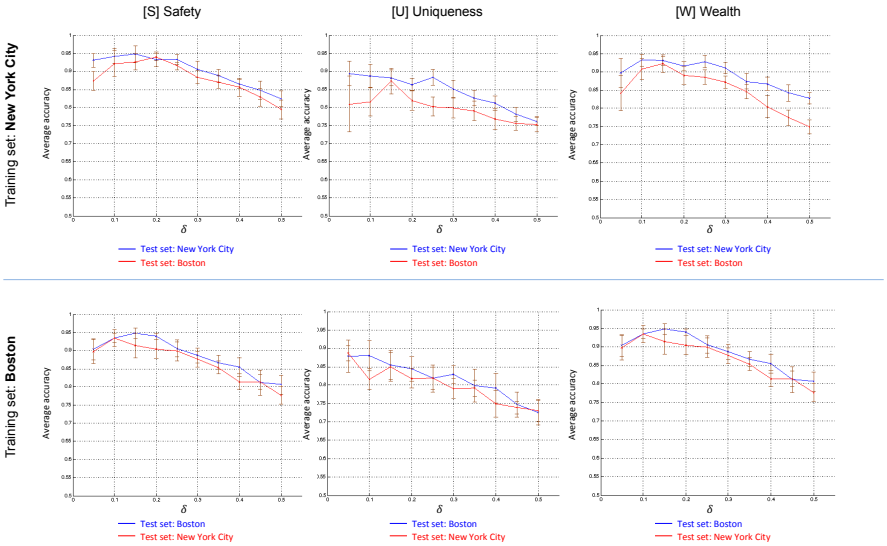


Fig. 4. Each figure shows the mean accuracy of the classification for different values of the δ parameter. The blue line represents performance reported on images from the same city as the training data. The red line represents the performance reported on images from a different city than those used for training.

We parameterize the classification problem by a variable δ and calculate performance as we adjust δ. As we move the value of our parameter δ the problem becomes more difficult since the visual appearance of the positive and negative images starts to become less evident up to the point when $\delta = 0.5$. At the same time when δ has smaller values the positive and negative images are easier to classify but we have access to less data.

We learn models to predict $y_{i,k}$ from input image representations x_i using an ℓ_2-regularized with a squared hinge-loss function linear SVM classifier:

$$\hat{y}_{i,k} = sgn(w_k^\intercal x_i) \tag{4}$$

$$w_k = \arg\min_{\overset{*}{w}_k} \frac{1}{2} \overset{*}{w}_k^\intercal \overset{*}{w}_k + c \sum_{i=1}^{n} (max(0, 1 - \breve{y}_{i,k} \overset{*}{w}_k^\intercal \breve{x}_i))^2 \tag{5}$$

Where we set the regularization parameter c using held-out data and learn w_k using training data $\{\breve{x}_i, \breve{y}_{i,k}\}$.

We examine two scenarios: a) training and testing perceptual prediction models on images from the same city, and b) training models on images from one city and testing on images from another city. We show some qualitative results of perceptual image classification in Figure 5.

We report classification performance on the Place Pulse dataset [28] in Figure 4 as mean average AUC, with error bars computed over 10 random splits for the SIFT + FV features. We performed the same analysis using Gist and DeCAF features and found them to be nearly on par for this task. Classification

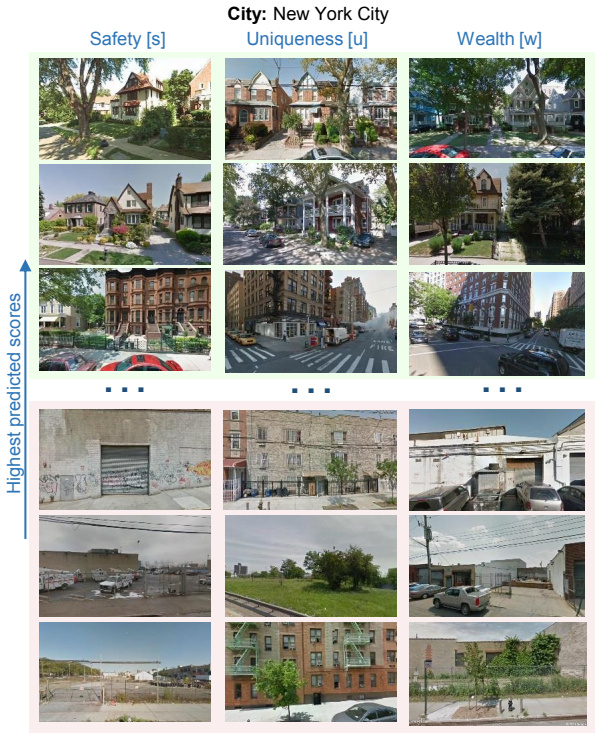


Fig. 5. Classification predictions on a our large densely sampled external dataset of street images of New York City. We show the images predicted as high and low – safety, uniqueness, wealthiness – at the top and the bottom respectively.

is evaluated for several values of δ, ranging from $\delta = 0.05$ to $\delta = 0.5$. The blue line in each plot represents accuracies for the scenario where we train and test on images from the same city. The red line in each plot represents accuracies for the scenario in which we train on one city and test on another city. For instance, the blue line in the top left plot in Figure 4 shows results for classifying images of New York City as {highly safe vs highly unsafe} using images of New York City as training data. The red line in the same plot corresponds to results of classifying images of Boston as {highly safe vs highly unsafe} using images of New York City as training data. The performance for training on one city and testing on another is slightly lower than training and testing on the same city, but reaches nearly the same performance for larger values of δ.

Several conclusions can be drawn from these plots. The first is that we can reliably predict the perceptual characteristics of safety, uniqueness, and wealth for streetview images. The second is that that uniqueness [U] seems to be the most difficult task to infer using our feature representations. This might be due to the more subjective definition of uniqueness.

We also find that we can train perceptual models on one city and then use them to reliably predict perceptions for another city, indicating the

Fig. 6. Regression predictions of safety for a previously unseen city. Here models are trained on images from Boston and New York City from the *Place Pulse v1.0* dataset and predictions are performed on a large newly collected streetview dataset of Chicago.

generalizability of our models to new places. This is important since our ultimate goal is to apply these methods to cities across the globe and collecting training data for every city in the world would be infeasible.

4.3 Regression

We also study perceptual characteristic prediction as a regression problem, where we want to predict aggregated human scores, defined in Eq. (1), using linear regression. Here, our ground truth labels are $y_{i,k} = q_{i,k}$ for image i and perceptual measure k. Therefore, we make predictions $\hat{y}_{i,k}$ as follows:

$$\hat{y}_{i,k} = w_k^\intercal x_i \tag{6}$$

$$w_k = \arg\min_{\overset{*}{w}_k} \frac{1}{2} \overset{*}{w}_k^\intercal \overset{*}{w}_k + c \sum_{i=1}^{n} (max(0, |\breve{y}_{i,k} - \overset{*}{w}_k^\intercal \breve{x}_i| - \epsilon))^2 \tag{7}$$

Where we optimize the squared loss error on the predictions subject to an ℓ_2 regularization on the parameters. We optimize for the regularization parameter c on held-out data and learn w_k using training data $\{\breve{x}_i, \breve{y}_{i,k}\}$.

Table 1. Results on the original Google Street View images from from the PlacePulse dataset (2011). We report the *Pearson product-moment correlation coefficient r* for the predicted regression values as compared to human perceptual scores for several training and testing data scenarios.

		Test on New York			Test on Boston		
Training data	Metric	Gist	FV	DeCaf	Gist	FV	DeCaf
New York	Safety	0.6365	**0.6869**	0.6808	0.6412	0.6566	**0.7008**
	Uniqueness	0.5265	0.5168	**0.5453**	0.4978	0.4358	**0.5186**
	Wealth	0.6149	0.6468	**0.6478**	0.5715	0.6001	**0.6608**
Boston	Safety	0.5972	0.6202	**0.6362**	0.6710	0.6740	**0.7180**
	Uniqueness	0.4474	0.3767	**0.4596**	0.5203	0.4941	**0.5471**
	Wealth	0.5640	0.5555	**0.6015**	0.5916	0.6419	**0.6782**

Table 2. Generalization of the PlacePulse annotations on updated Google StreetView images (2013). We report the *Pearson product-moment correlation coefficient r* for the predicted regression values as compared to human perceptual scores for several training and testing data scenarios.

		Test on New York			Test on Boston		
Training data	Metric	Gist	FV	DeCaf	Gist	FV	DeCaf
New York	Safety	0.5436	**0.5890**	0.5603	0.5165	0.5275	**0.5578**
	Uniqueness	0.4388	**0.4510**	0.4449	0.4072	0.3598	**0.4363**
	Wealth	0.5328	**0.5659**	0.5518	0.4698	0.4949	**0.5631**
Boston	Safety	0.5062	0.4895	**0.5211**	0.5531	**0.5839**	0.5757
	Uniqueness	0.4023	0.3479	**0.4158**	0.4208	0.3712	**0.4527**
	Wealth	0.4972	0.4801	**0.5173**	0.5238	0.5367	**0.5863**

Regression results for predicting safety, uniqueness, and wealth are presented in Table 1, computed over 10 folds of the data for Gist, SIFT-FV and DeCAF image descriptors. As in our classification experiments, we examine two scenarios: training and testing on the same city, and training on one city and testing on a different city. We find that our models are able to predict perceptual scores well, with r-correlation coefficients ranging from 0.4 to 0.7. Again we find uniqueness to be the most challenging perceptual characteristic for prediction. Here DeCAF features provided the highest generalization performance when testing on data from a different city. We show some qualitative results across the spectrum of predicted scores for the city of Chicago in Figure 6 (Note we did not have images of Chicago available for training). We additionally show prediction scores for several metrics on a map in Figure 8.

Generalization across Time: The original PlacePulse dataset annotations were collected in 2011 with the available Google Street View images at that time.

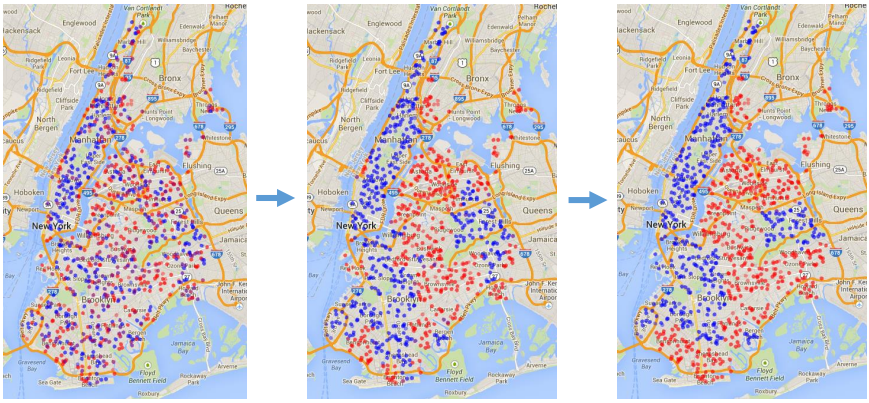


Fig. 7. The input map on the left are isolated predictions of perceptual safety for New York City. The next two images are joint predictions of safety/unsafety using our collective model with different smoothing parameters.

We additionally downloaded updated images for the same locations and views, most of which were taken in 2013. We run the same regression experiments on this set of images using the original perceptual scores as labels and show results in Table 2. We find that even though performance drops somewhat we are still able to learn representative and reasonably accurate models for each concept.

5 Collective Urban Perception

In the previous models, prediction for classification and regression was performed independently for each image. However, images of a place are not independent. The safety of one city block is tightly correlated with the safety of the next block down the street. In this section, we explore models for collective inference of perceptual characteristics within a city. In particular, we model a city as a graph where each node $n_i \in N$ is represented by a set of variables $\{p_i = (lat_i, lon_i), x_i\}$ where p_i is a latitude-longitude coordinate and x_i is the feature representation for the image. We connect the nodes in the graph to define the edge set E by associating each node n_i with its closest K neighbors based on the euclidean distance between pairs of node coordinates (p_i, p_j). For our experiments we use a connectivity factor of $K = 10$.

Now, let's say our goal is to label every node in the graph as unsafe or not. We first define unsafe as any point in our training data that has an image with a perceptual score $q_{i,s}$ in the bottom 25% of the training set. We set our goal to predict a joint labeling $\hat{Y} = \{y_i\}$ that maximizes:

$$\hat{Y} = \arg\max_{Y} \prod_{i} \Phi_1(y_i|x_i, w_s) \prod_{i,j \in E} \Phi_2(y_i, y_j|x_i, x_j, p_i, p_j, \alpha_1, \alpha_2) \tag{8}$$

$$-\ln \Phi_1 = y_i w_s^{\intercal} x_i \tag{9}$$

$$-\ln \Phi_2 = \left(\frac{\alpha_1}{\|x_i - x_j\|} + \frac{\alpha_2}{\|p_i - p_j\|} \right) \cdot 1[y_i = y_j] \tag{10}$$

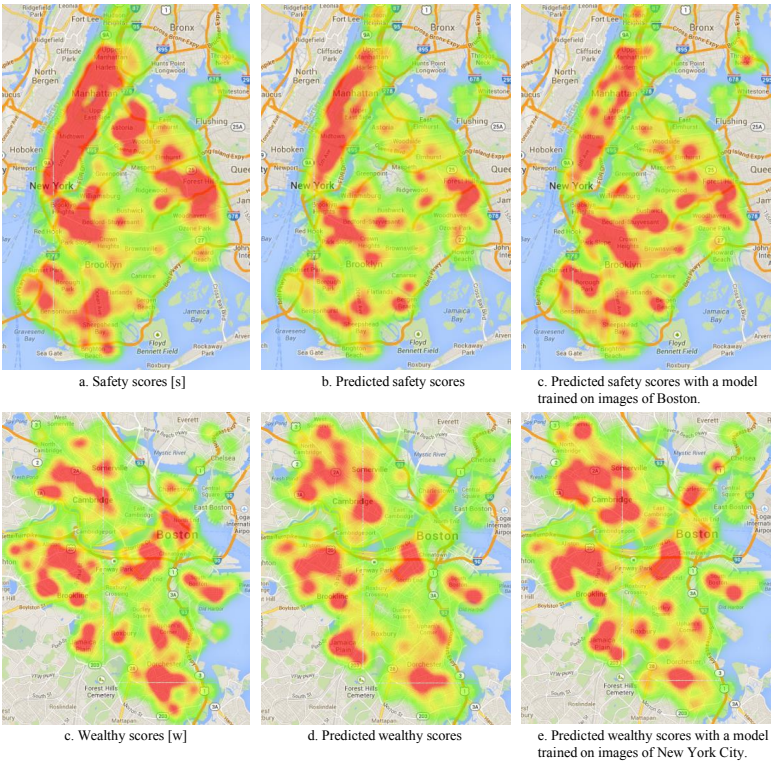


Fig. 8. Regression results scaled and shown as a heatmap for all the point locations in the Place Pulse Dataset. Left column shows ground truth scores, middle column shows predictions from the regression model, and right column shows predictions of the regression model when trained on a different city.

Where the unary potentials parameter w_s is based on our regression model (Section 4.3). The pairwise potentials for smoothing are based on two criteria: Visually similar images should be encouraged to take the same label, and images that are spatially close to each each other should be encouraged to take the same label. This global optimization is in general difficult, but because we are using submodular potentials we can optimize this in polynomial time using Graphcuts [3].

We use this model to jointly predict perceptual scores for least safe, unique, and wealthy images coherently across all images in New York City. Results for average f1-scores computed over 10 folds (with line search to tune parameters α_1 and α_2) are shown in Table 3 for both the SIFT-FV and DeCAF features. To reduce correlations between images used in training and testing we select train-test splits of the data by clustering the data points using k-means on image coordinates ($k = 10$). Each cluster is used as the test data for one fold and the rest of the images are used for training.

Table 3. F1-scores for predicting perceptions of least safe, least unique and least wealthy places using isolated predictions and our collective unsafety prediction model

		¬Safe	¬Unique	¬Wealth
Isolated prediction	[SIFT + FV]	0.6077	0.4420	**0.5755**
Isolated prediction	[DeCAF]	0.5929	0.4652	0.5613
Collective prediction	[SIFT + FV]	0.6069	0.4457	0.5700
Collective prediction	[DeCAF]	**0.6089**	**0.4777**	0.5545

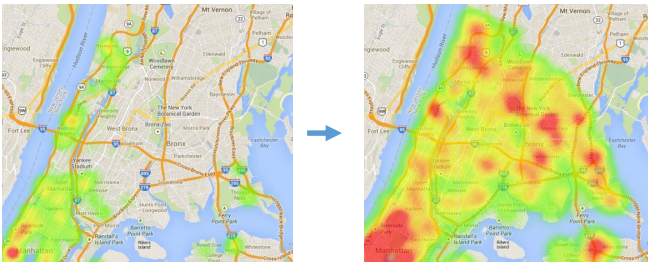


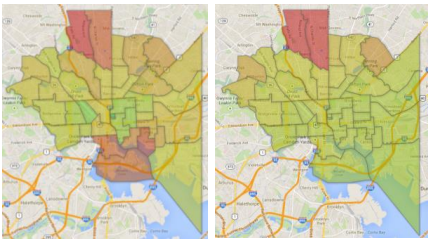
Fig. 9. Large scale experiments: Left map represents predictions on the original Place Pulse dataset for New York, Right map shows the result of applying classification models to our more densely and broadly sampled data

We find a positive improvement for predicting the ground-truth binary labels jointly rather than independently for predictions of least safe and least unique places. For predicting which images are not wealthy we don't find any improvement, perhaps indicating that wealthiness is more localized.

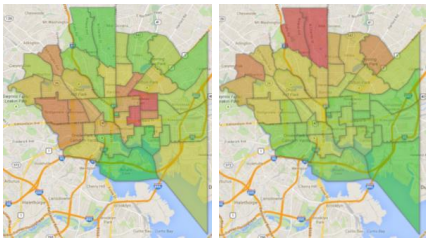
On the qualitative side we now have, akin to foreground-background segmentation, a model that can produce arbitrarily dense or sparse region representations of safe/unsafe areas depending on the parameter choice of the pairwise potentials. We show some results to this effect in Figure 7. From these maps, we can see a birds eye view of which parts of New York City are most safe or unsafe. If we use less smoothing we can see more fine-grained predictions of safety at the neighborhood level. Notably the blue area includes Manhattan and certain neighborhoods in Brooklyn and Queens like Park Slope and Forest Hills which are known to be particularly safe areas of these boroughs.

6 Additional Experiments and Results

Large Scale Experiments on Unlabeled Data: So far, we have been evaluating our models on the *Place Pulse v1.0* dataset, but we have also collected a much larger, densely sampled dataset of the original two cities (New York City and Boston) and two new cities (Baltimore and Chicago). Therefore, we run our models on these datasets as well. In Figure 9 we show predictions on our New York City dataset compared to the original samples from Place Pulse. Our dataset contains not only denser sampling, but also areas that were not present



a. The map on the left shows household income statistics, the map on the right shows our predicted scores.



b. The map on the left shows homicide statistics, the map on the right shows our predicted safety scores.

Fig. 10. The pair of maps on the left showcase the positive correlation between household income statistics and our predicted perceptual scores of wealthiness. The pair of maps on the right showcase the negative correlation between homicide statistics and our predicted perceptual scores of safety.

in the original study. For instance we include samples from extended areas like the Bronx. Figure 9 shows qualitative results for perceptions of wealth for the Bronx using our predicted scores. The results seem to confirm anecdotal evidence of affluence of certain areas in the Bronx such as Riverdale or Country Club, both upper middle class neighborhoods[1].

Correlation of Our Models with Crime Statistics: The authors of the Place Pulse dataset found that human perception judgments are informative about crime statistics of homicides for New York City. We go further, predicting safety, wealth, and uniqueness on two cities for which we have no ground truth perceptual judgments. We compute correlations between our predictions and reported statistics of homicides and household income per county[2]. We aggregate our predictions over counties and compare to reported statistics in Figure 10. We find a moderate positive Pearson-correlation coefficient of 0.51 between Baltimore household income and our predictions of wealth. In Figure 10a we observe good predictions for the two wealthiest counties in Baltimore, but miss a third cluster in South Baltimore. We also find a moderate negative Pearson-correlation coefficient of -0.36 between homicide statistics and our predictions of safety (Figure 10b). If we restrict our analysis to counties for which we have a larger number of sample images n then we obtain stronger correlations: [0.53 ($n > 200$), 0.61 ($n > 300$) for income/wealth predictions and [-0.41 ($n > 200$),-0.47 ($n > 300$)] for crime/safety predictions (by even denser sampling we could potentially extend this to all locations). For Chicago we find weaker correlation coefficients of 0.32 for wealth and -0.21 for safety when compared to similar statistics.

[1] `http://en.wikipedia.org/wiki/Riverdale,_Bronx` and `http://en.wikipedia.org/wiki/Country_Club,_Bronx`

[2] Data obtained from the Baltimore City Health Department 2011 report and from `http://www.robparal.com/` for Chicago.

7 Conclusions

In this paper we have shown that visual models can predict human perceptions of place. In particular, we demonstrated experimental evaluations for classification and regression predictions of safety, uniqueness, and wealth. We also produced models for joint prediction of perceptual characteristics. Finally, we demonstrated uses of our model for predicting perceptual characteristics at city scale and confirmed our findings for novel cities through correlations with crime statistics. These findings take us one step toward understanding sense of place.

Acknowledgments. This work was funded in part by NSF Awards 1445409 and 1444234.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM 54(10), 105–112 (2011)
2. Arietta, S., Efros, A., Ramamoorthi, R., Agrawala, M.: City forensics: Using visual elements to predict non-visual city attributes. IEEE Transactions on Visualization and Computer Graphics (2014)
3. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9), 1124–1137 (2004)
4. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC 2011 (2011)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
6. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1657–1664. IEEE (2011)
7. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes paris look like paris? ACM Transactions on Graphics (SIGGRAPH) 31(4) (2012)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. ArXiv e-prints (October 2013)
9. Frahm, J.-M., et al.: Building rome on a cloudless day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 368–381. Springer, Heidelberg (2010)
10. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: ICCV (2009)
11. Hays, J., Efros, A.A.: Im2gps: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
12. Isola, P., Parikh, D., Torralba, A., Oliva, A.: Understanding the intrinsic memorability of images. In: NIPS, pp. 2429–2437 (2011)

13. Juneja, M., Vedaldi, A., Jawahar, C.V., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
14. Khosla, A., An, B., Lim, J.J., Torralba, A.: Looking beyond the visible scene. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Ohio, USA (June 2014)
15. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.: Associative hierarchical crfs for object class image segmentation. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 739–746. IEEE (2009)
16. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE (2006)
17. Lee, Y.J., Efros, A.A., Hebert, M.: Style-aware mid-level representation for discovering visual connections in space and time. In: 2011 IEEE International Conference on Computer Vision (ICCV) (2013)
18. Li, L.-J., Su, H., Lim, Y., Fei-Fei, L.: Objects as attributes for scene classification. In: Kutulakos, K.N. (ed.) ECCV 2010 Workshops, Part I. LNCS, vol. 6553, pp. 57–69. Springer, Heidelberg (2012)
19. Lynch, K.: The image of the city, vol. 11. MIT Press (1960)
20. Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1784–1791. IEEE (2011)
21. Naik, N., Philipoom, J., Raskar, R., Hidalgo, C.: Streetscore-predicting the perceived safety of one million streetscapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 779–785 (2014)
22. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
23. Park, J., Newman, M.E.: A network-based ranking system for us college football. Journal of Statistical Mechanics: Theory and Experiment 2005(10), P10014 (2005)
24. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2751–2758. IEEE (2012)
25. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
26. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 413–420. IEEE (2009)
27. Quercia, D., O'Hare, N.K., Cramer, H.: Aesthetic capital: What makes london look beautiful, quiet, and happy? In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2014, pp. 945–955. ACM, New York (2014), http://doi.acm.org/10.1145/2531602.2531613
28. Salesses, P., Schechtner, K., Hidalgo, C.A.: The collaborative image of the city: mapping the inequality of urban perception. PloS One 8(7), e68400 (2013)
29. Sudderth, E., Torralba, A., Freeman, W., Willsky, A.: Learning hierarchical models of scenes, objects, and parts. In: ICCV (October 2005)
30. Tighe, J., Lazebnik, S.: Finding things: Image parsing with regions and per-exemplar detectors. In: CVPR (2013)

31. Tuan, Y.F.: Landscapes of fear. Basil Blackwell, Oxford (1980)
32. Wang, H., Gould, S., Koller, D.: Discriminative learning with latent variables for cluttered indoor scene understanding. Communications of the ACM, Research Highlights 56, 92–99 (2013)
33. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492. IEEE (2010)
34. Zamir, A.R., Shah, M.: Accurate image localization based on google maps street view. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 255–268. Springer, Heidelberg (2010)