Learning High-level Judgements of Urban Perception

Presented by
Leandra Irvine
Yuanyang (Ian) Zheng

Task

Perceptual characteristics of city images

1) Safety
2) Uniqueness
3) Wealth

Data

1) Place Pulse 1.0 -> Pairwise comparison -> Perceptual score (NYC & Boston)

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

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

2) Unscored sampled data from Google Streetview API (NYC, Boston, Chicago & Baltimore)

Data representation

- Gist
- SIFT (Scale-invariant feature transform) + FV
- DeCAF (Deep Convolutional Activation Feature)

```
nn.Sequential {
  [input -> (1) -> (2) -> output]
  (1): nn.Sequential {
    [input -> (1) -> (2) -> (3) -> (4) -> (5) -> (6) -> (7) -> (8) -> (9) -> (10) -> (11) -> (12) -> (13) -> (14) -> (1
    (1): nn.SpatialConvolution(3 -> 64, 11x11, 4,4, 2,2)
    (2): nn.SpatialBatchNormalization
    (3): nn.ReLU
    (4): nn.SpatialMaxPooling(3x3, 2,2)
    (5): nn.SpatialConvolution(64 -> 192, 5x5, 1,1, 2,2)
    (6): nn.SpatialBatchNormalization
    (7): nn.ReLU
    (8): nn.SpatialMaxPooling(3x3, 2,2)
    (9): nn.SpatialConvolution(192 -> 384, 3x3, 1,1, 1,1)
    (10): nn.SpatialBatchNormalization
    (11): nn.ReLU
    (12): nn.SpatialConvolution(384 -> 256, 3x3, 1,1, 1,1)
    (13): nn.SpatialBatchNormalization
    (14): nn.ReLU
    (15): nn.SpatialConvolution(256 -> 256, 3x3, 1,1, 1,1)
    (16): nn.SpatialBatchNormalization
    (17): nn.ReLU
    (18): nn.SpatialMaxPooling(3x3, 2,2)
  }
  (2): nn.Sequential {
    [input -> (1) -> (2) -> (3) -> (4) -> (5) -> (6) -> (7) -> (8) -> (9) -> (10) -> (11) -> output]
    (1): nn.View(9216)
    (2): nn.Dropout(0.500000)
    (3): nn.Linear(9216 -> 4096)
    (4): nn.BatchNormalization
    (5): nn.ReLU
    (6): nn.Dropout(0.500000)
    (7): nn.Linear(4096 -> 4096)
    (8): nn.BatchNormalization
    (9): nn.ReLU
    (10): nn.Linear(4096 -> 1000)
    (11): nn.LogSoftMax
  }
}
```

Classification

Two classes:

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

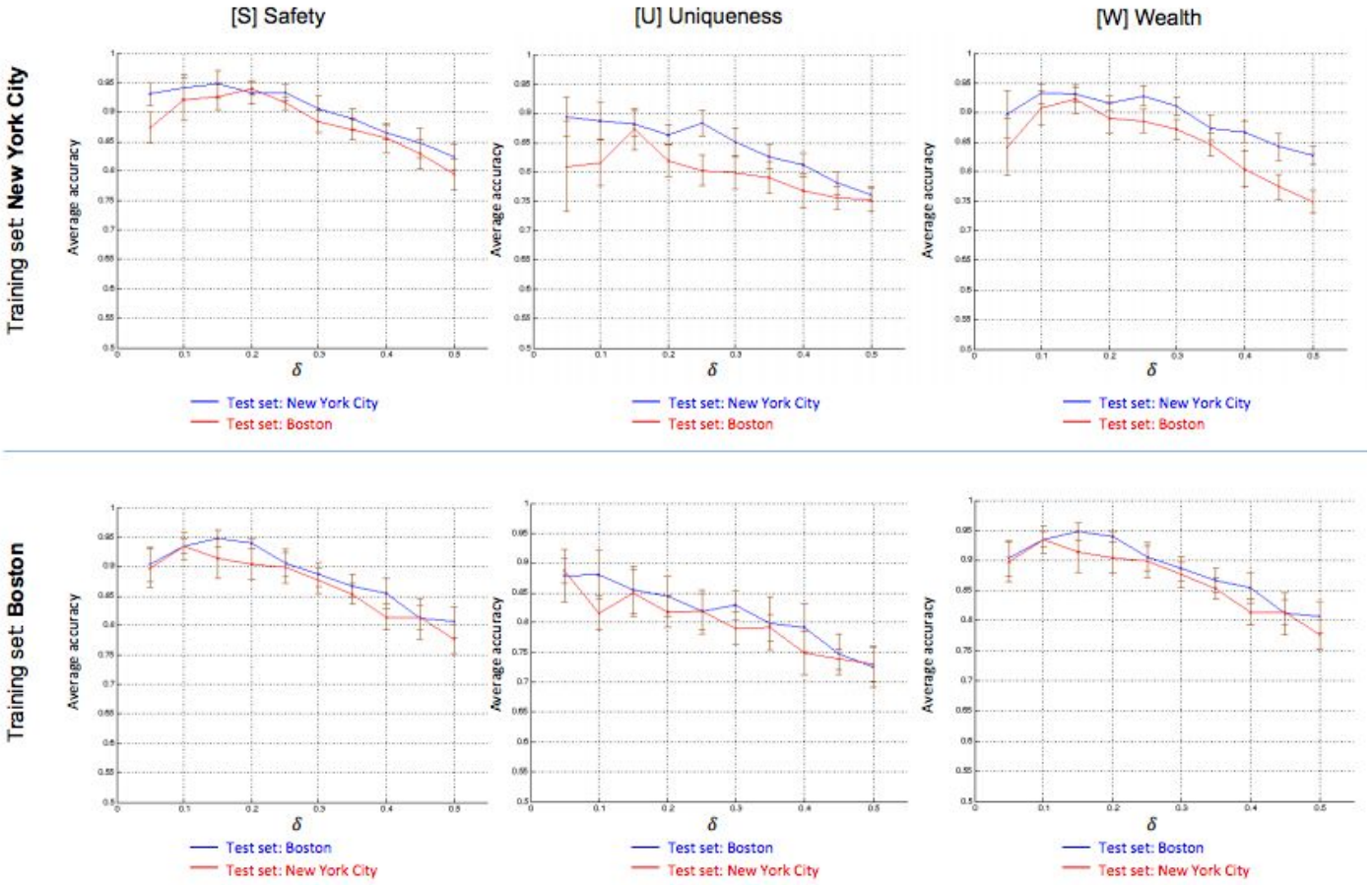
Linear SVM with L2 regularization and squared hinge loss function

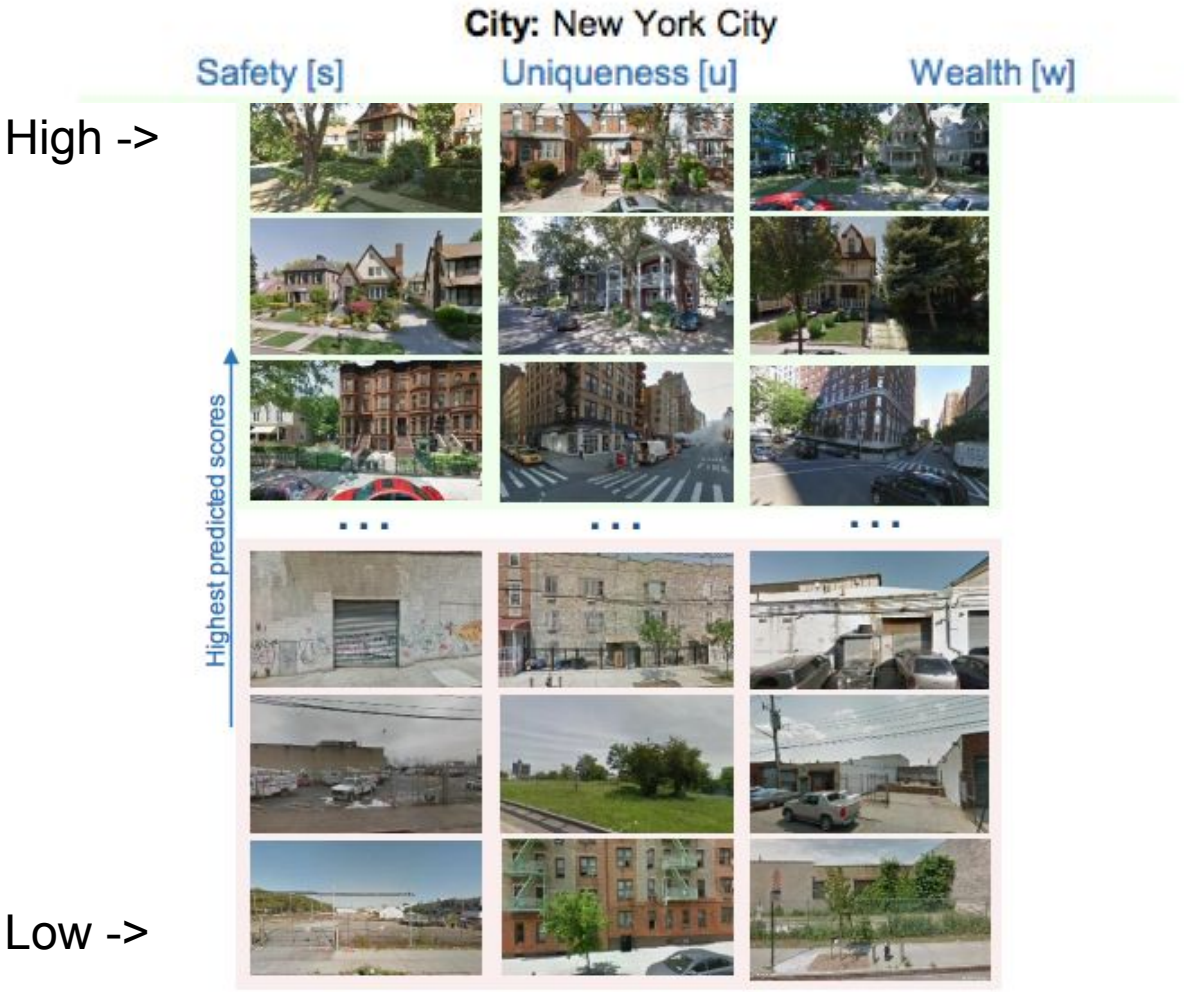
$$\hat{y}_{i,k} = sgn(w_k^\intercal x_i)$$

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

Results

1. Predictability
2. Uniqueness
   less accurate
3. Generalization




City: New York City
Safety [s]
Uniqueness [u]
Wealth [w]
High ->
Highest predicted scores
Low ->

Regression

Linear with L2 regularization

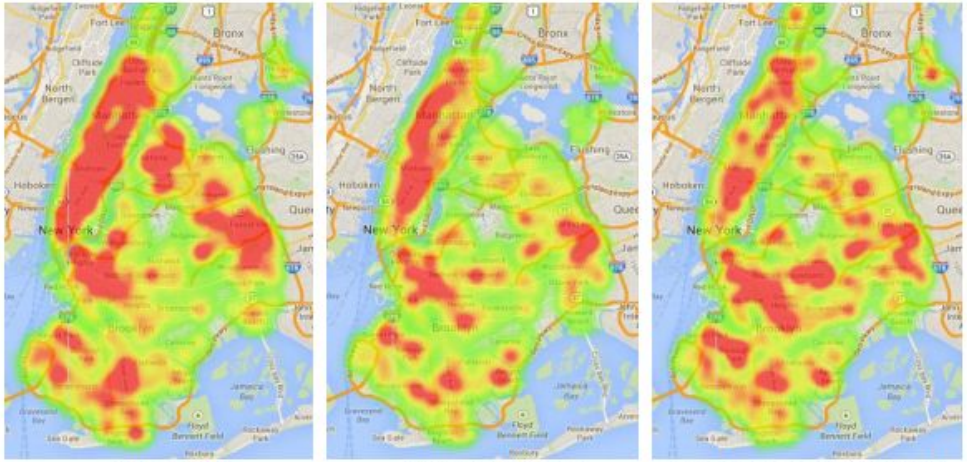
$$\hat{y}_{i,k} = w_k^{\intercal} x_i \tag{6}$$

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

Results

1. Generalization across time
   a. 2011 and 2013

NYC ->

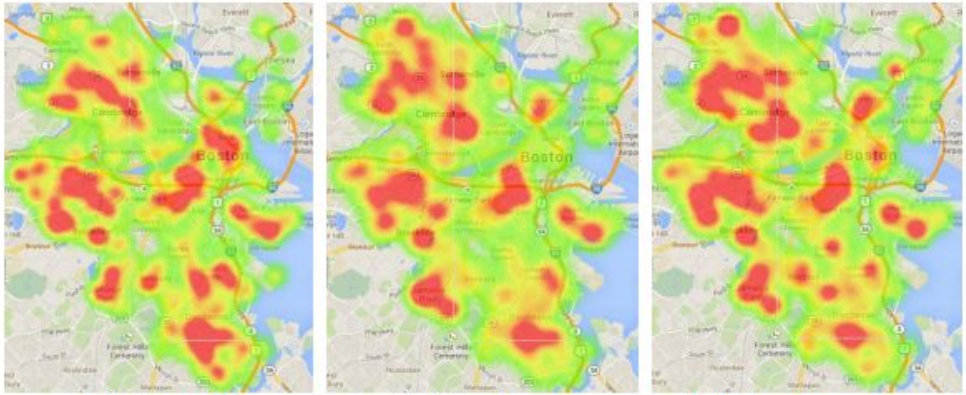


a. Safety scores [s]

b. Predicted safety scores

c. Predicted safety scores with a model trained on images of Boston.

Boston ->



c. Wealthy scores [w]

d. Predicted wealthy scores

e. Predicted wealthy scores with a model trained on images of New York City.

Collective Urban prediction

Graph approach K = 10, Euclidean distance

Visually similar images should be encouraged to take the same label, and images that are **spatially close to each other** should be encouraged to take the same label.

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

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

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

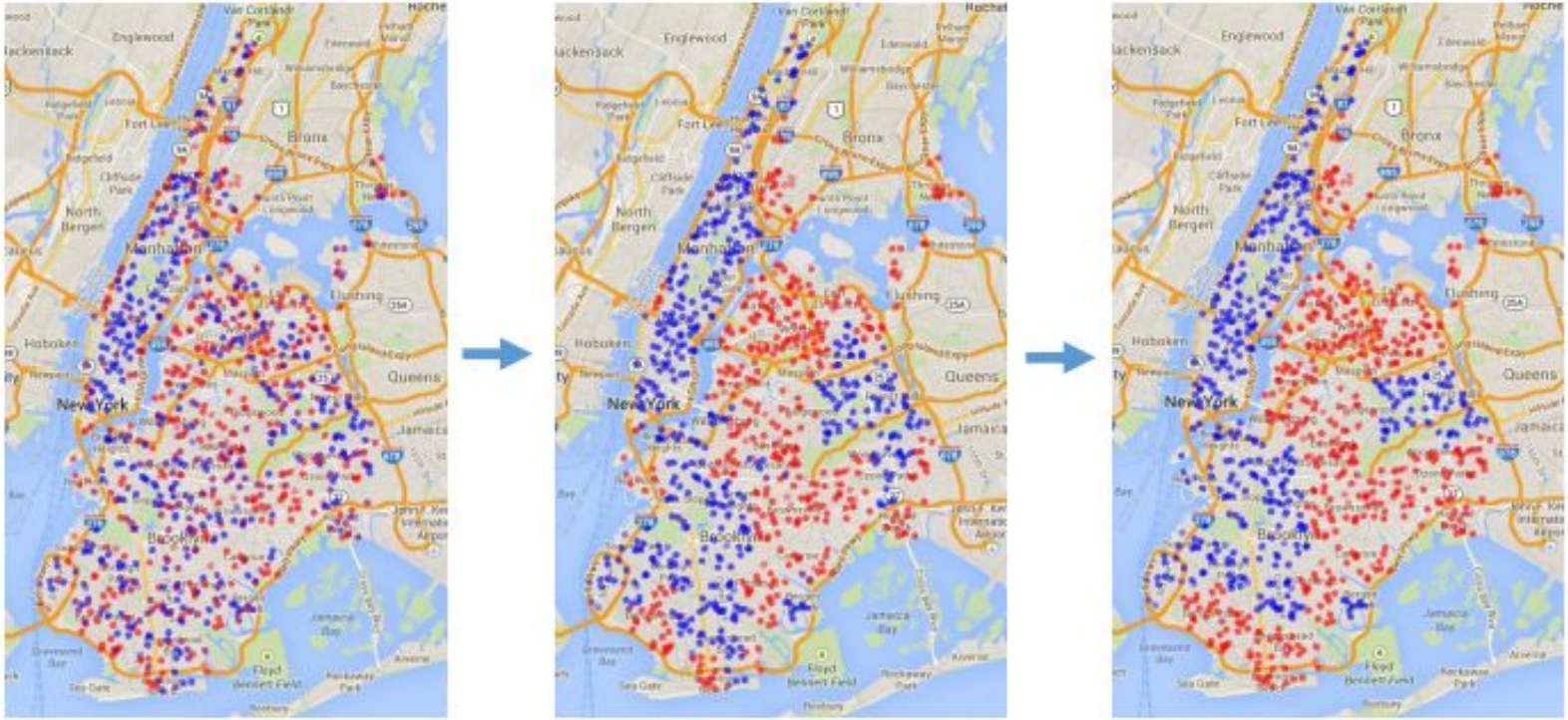


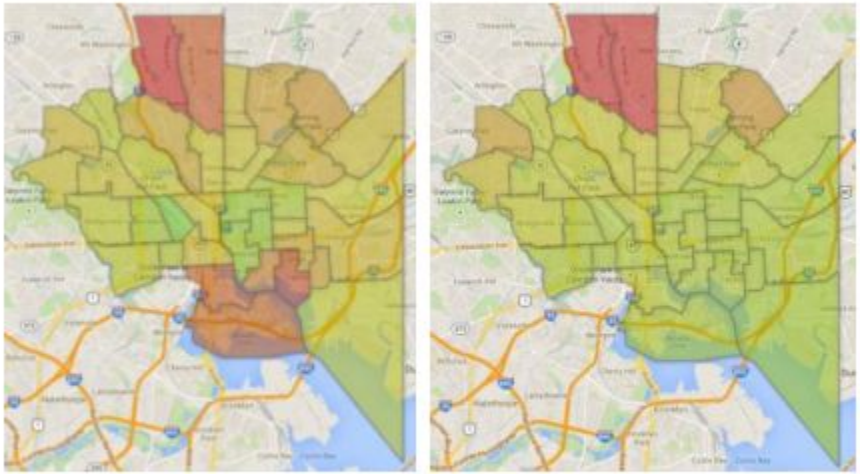
Fig. 7. The input map on the left are isolated predictions of perceptual safety for New York City. The next two images are joint predictions of safety/unsafety using our collective model with different smoothing parameters.

Continuous Regions

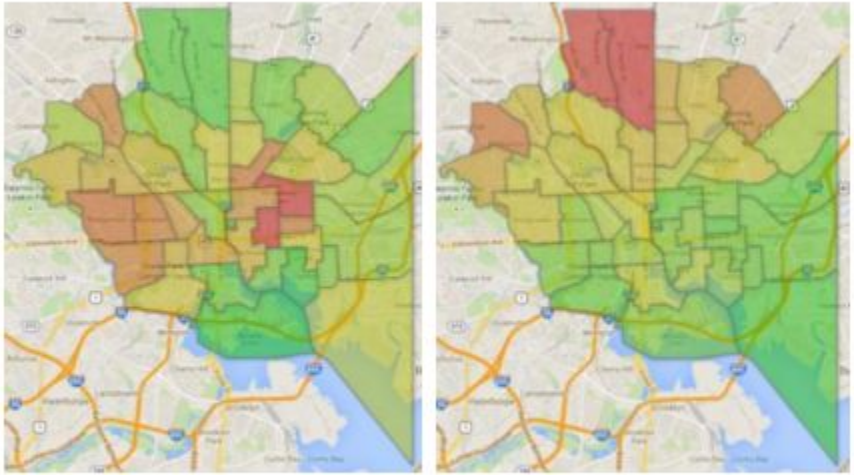
Additional Validation

NYC using the Streetview API data

Baltimore using Income/Crime Statistics



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.

More Recent Work (Aug 2016)

-Larger Scale

-Siamese Network

Deep Learning the City: Quantifying Urban Perception At A Global Scale

Abhimanyu Dubey[1], Nikhil Naik[3], Devi Parikh[2]
Ramesh Raskar[3], César A. Hidalgo[3]

[1] Indian Institute of Technology Delhi
abhimanyu1401@gmail.com
[2] Virginia Tech
parikh@vt.edu
[3] MIT Media Lab
{naik,raskar,hidalgo}@mit.edu

Abstract. Computer vision methods that quantify the perception of urban environment are increasingly being used to study the relationship between a city's physical appearance and the behavior and health of its residents. Yet, the throughput of current methods is too limited to quantify the perception of cities across the world. To tackle this challenge, we introduce a new crowdsourced dataset containing 110,988 images from

Discussion