STANFORD UNIVERSITY

Using **location** to **see**!

Amir R. Zamir

Phone!

Phone + Camera!

Phone + Camera + GPS!

Phone + Camera + GPS!

Vision + Location!

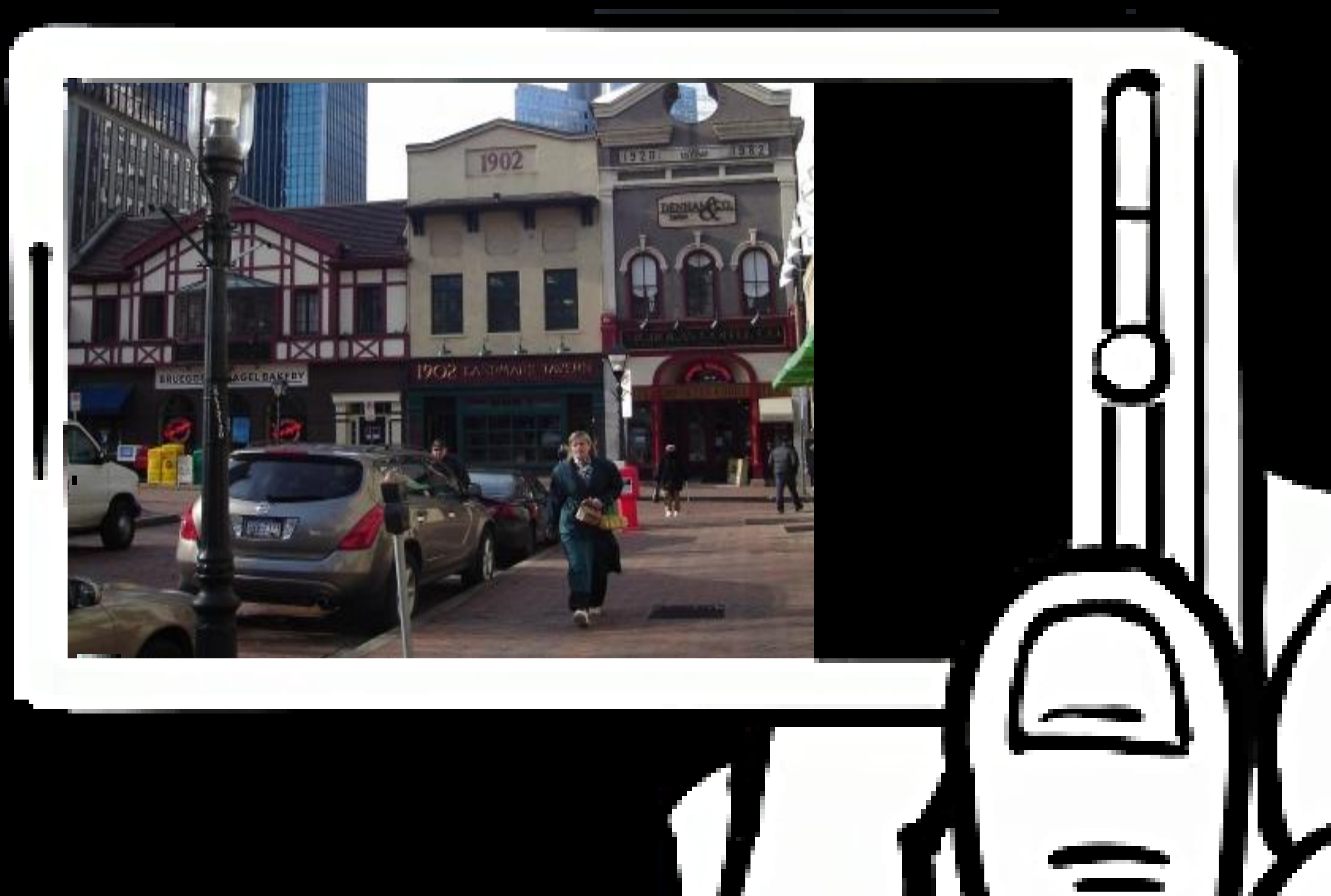
Why not just camera?
Why not just GPS?

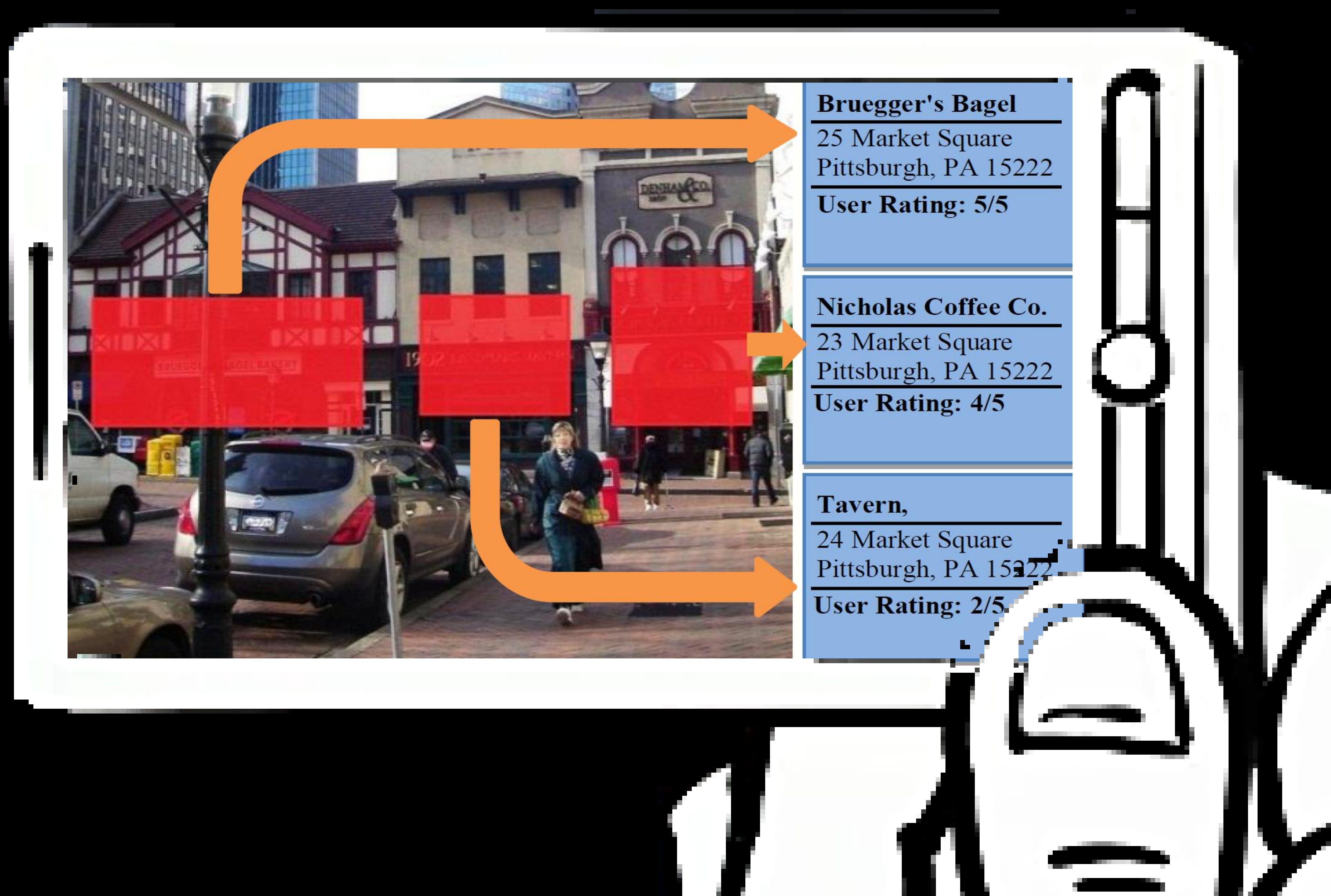
Vision + Location!


1902
1920
1982
DENHAM & CO.
salon
NICHOLAS COFFEE CO.
1902 LANDMARK TAVERN
COFFEE ROASTERS
IMPORTED FOODS
BAGEL BAKERY
SL13130
GNB-7333


1902
1920
1982
DENHAM & CO.
salon
NICHOLAS COFFEE CO.
1902 LANDMARK TAVERN
IMPORTED FOODS
COFFEE ROASTERS
BAGEL BAKERY
SL13130
GNB-7333

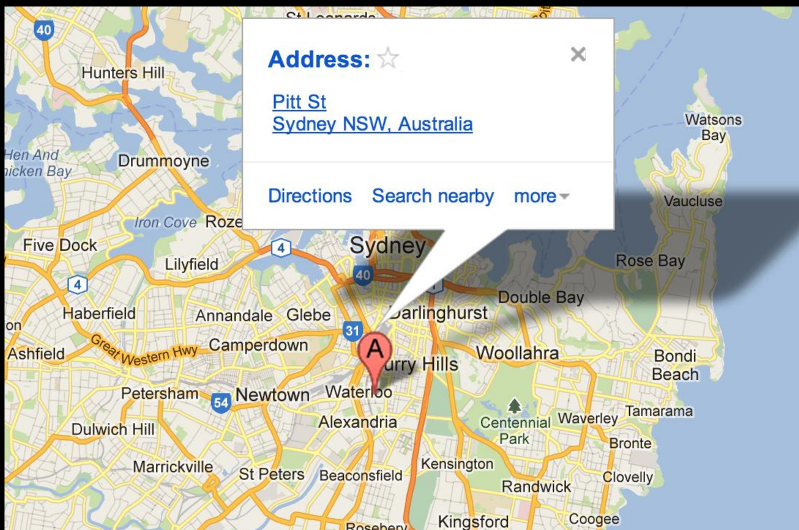

1902
1920
1982
DENHAM & CO.
salon
1902
SL13130

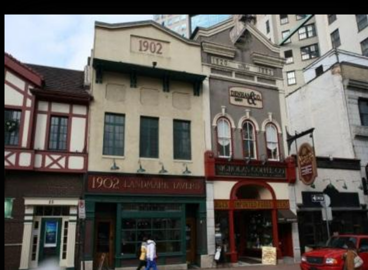

1902
1902 LANDMARK TAVERN



Bruegger's Bagel
25 Market Square
Pittsburgh, PA 15222
User Rating: 5/5
Nicholas Coffee Co.
23 Market Square
Pittsburgh, PA 15222
User Rating: 4/5
Tavern,
24 Market Square
Pittsburgh, PA 15222
User Rating: 2/5

Geo-localization

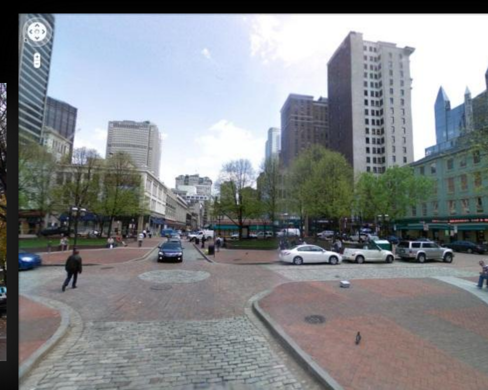




Query

Match – Error: 7.6 m

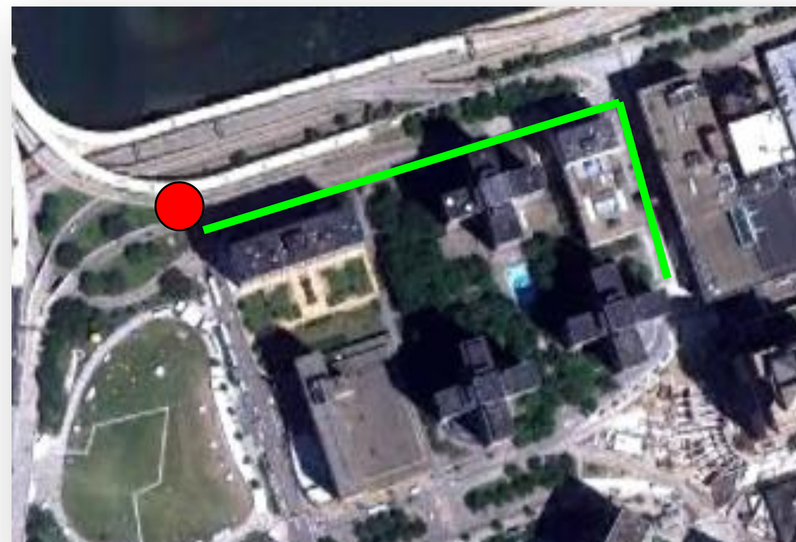
Query



Match – Error: 6.9 m

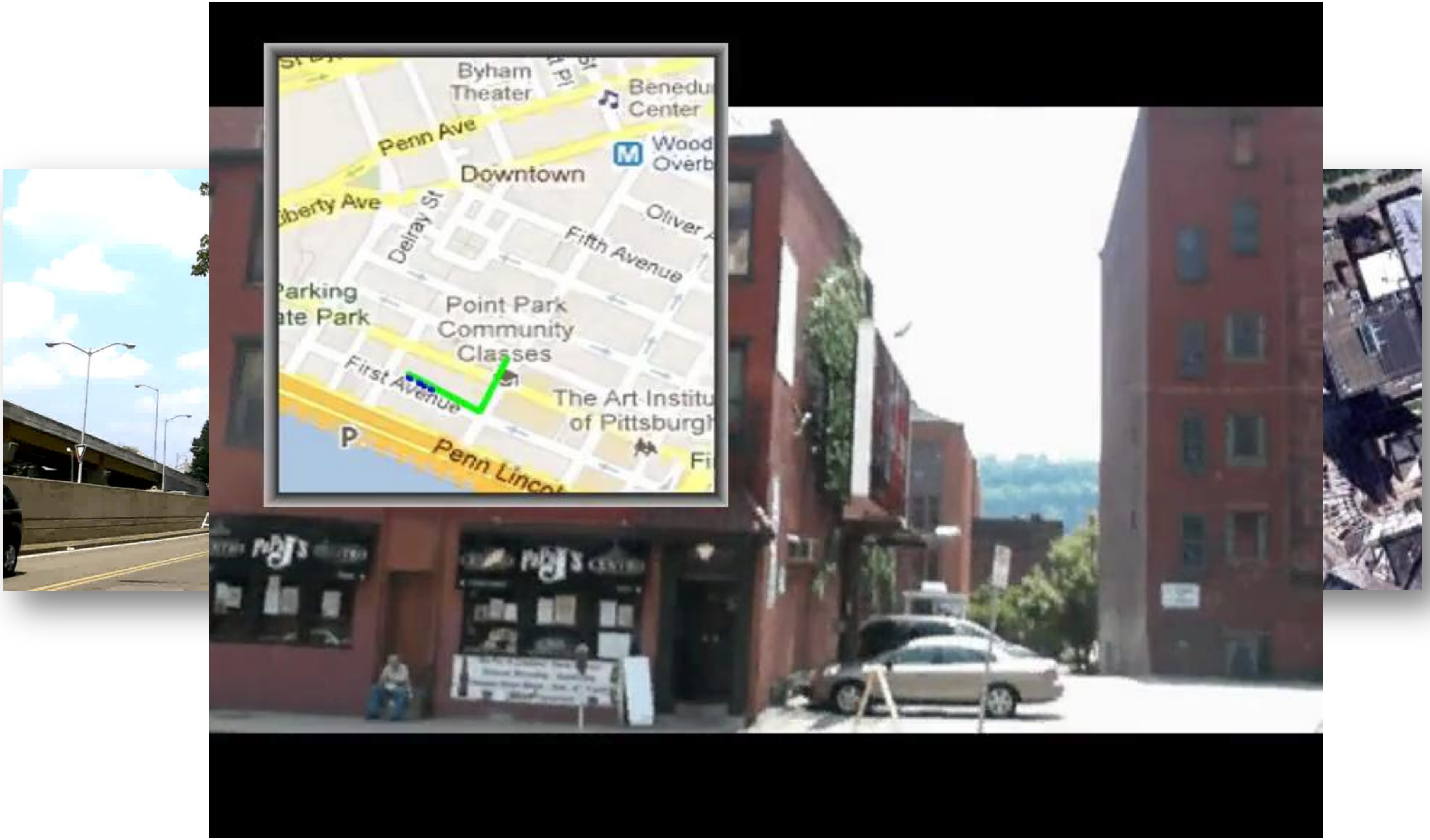
Where the user have visited?

Input Video



Desired output

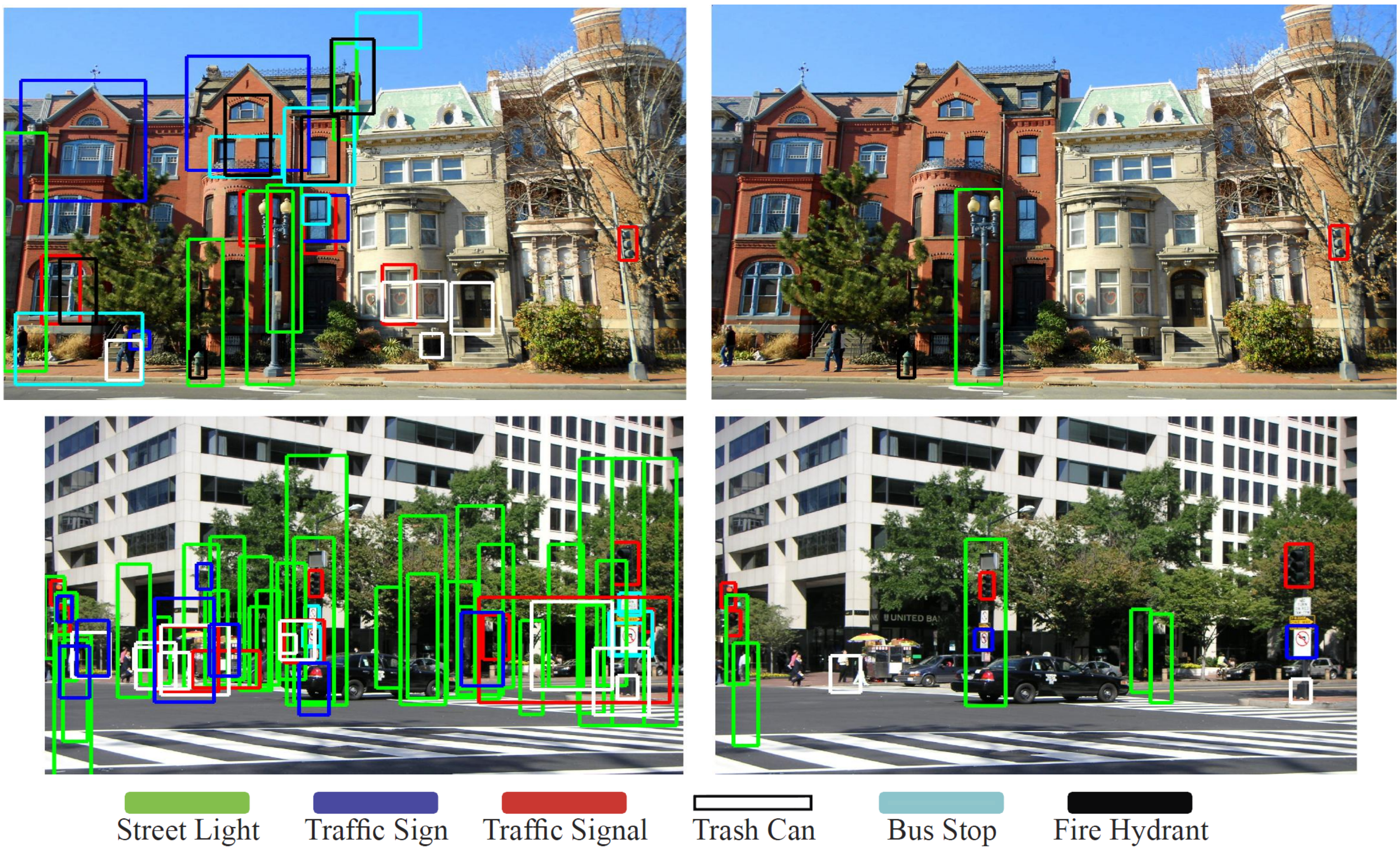
Where the user have visited?



Object Detection

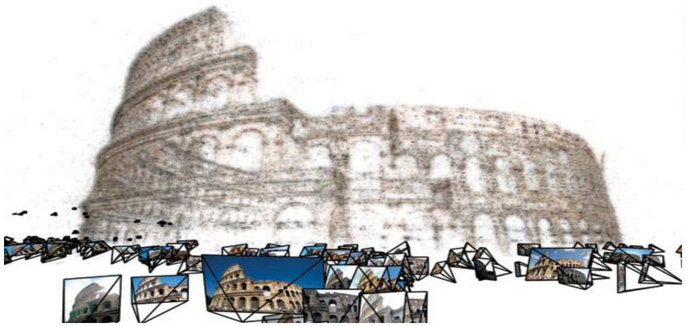
Conventional Object Detection

Location-assisted Object Detection, Smartphone Image

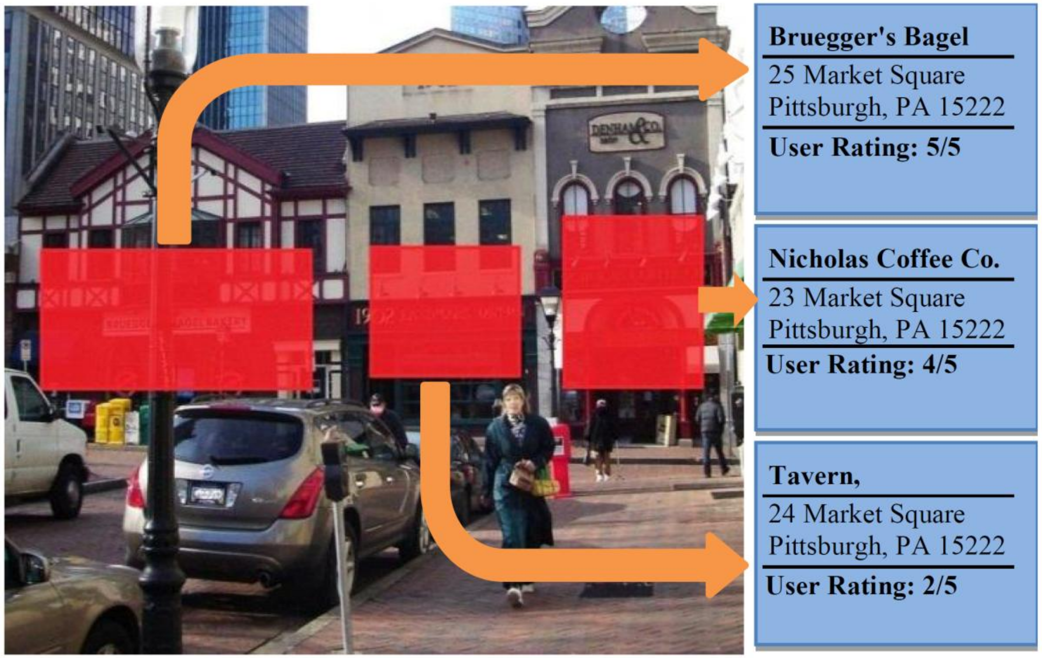


Not limited to mobile devices

Geographical Data Organization



Building 3D models [59]



Assisted Content Analysis on Mobile Devices

Geo-spatial Perspective

- Three fundamental questions in **Geo-spatial Analysis of Ground Level Images and Videos**:
  1) How to automatically geo-localize images and videos?
  2) How to refine the geo-location of already geo-tagged data?
  3) How to utilize the geo-location in content analysis?

- **Automatic Geo-localization:**
  - Image geo-localization using Generalized Graphs
  - Video Geo-localization and trajectory extraction
- **Robust Refinement of geo-location using Random Walks**
- **Location-Aware image understating:**
  - Location-aware object detection
  - Precise recognition of storefronts in images

Paper:

Image Geo-localization Based on Multiple Nearest Neighbor Feature Matching Using Generalized Graphs. In ***T-PAMI***, 2014.

“Where Am I?”

- Problem:

  Accurate Image Localization

Input

Mere Visual Information (Images)

Output

Location in Terms of λ (Lon.) and φ (Lat.)

“Where Am I?”

- Problem:

  Accurate Image Localization

Input

Mere Visual Information (Images)

Output

Location in Terms of λ (Lon.) and φ (Lat.)

φ=40.4419, λ=-79.9986

Google Maps Street View Dataset

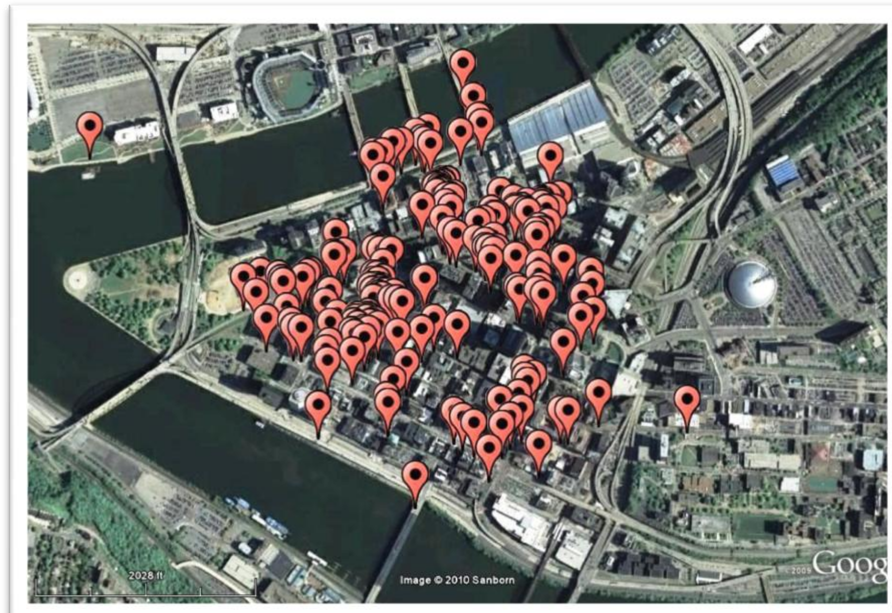
332 6th Ave, Pittsburgh, PA

Google Maps Street View Dataset

332 6th Ave, Pittsburgh, PA $\varphi=40.4418, \lambda=-79.9987$

Street View Dataset

Reference Images Place Marks



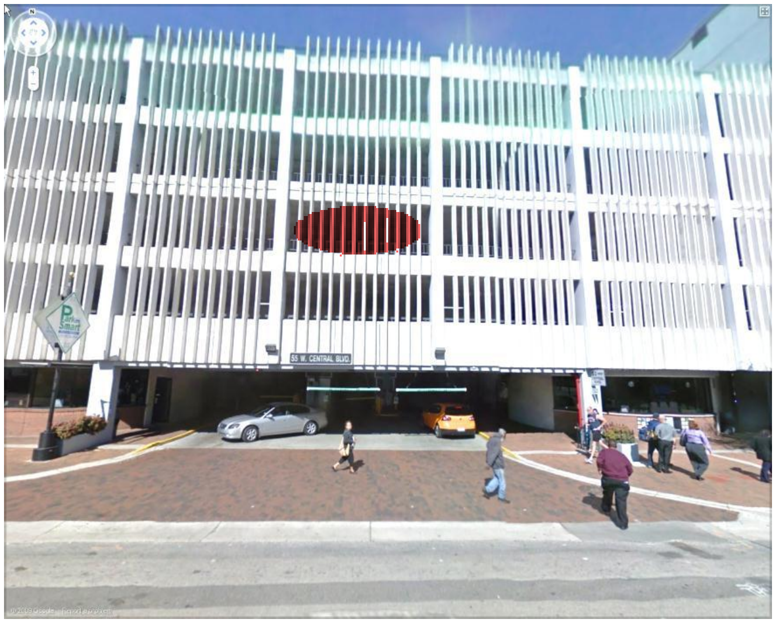
Query Images

Pittsburgh, PA

Ambiguity of local features

=

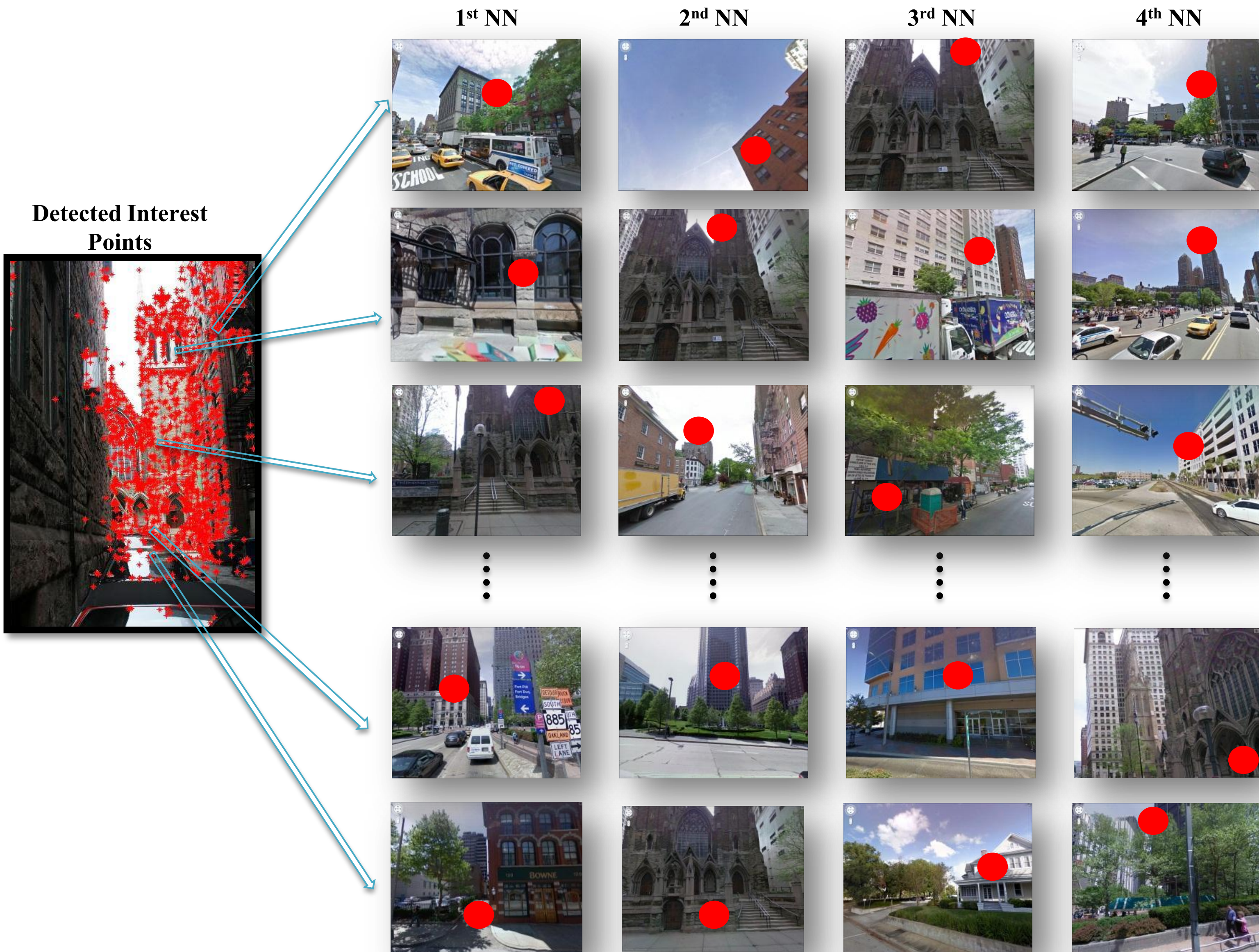
Ambiguity of local features



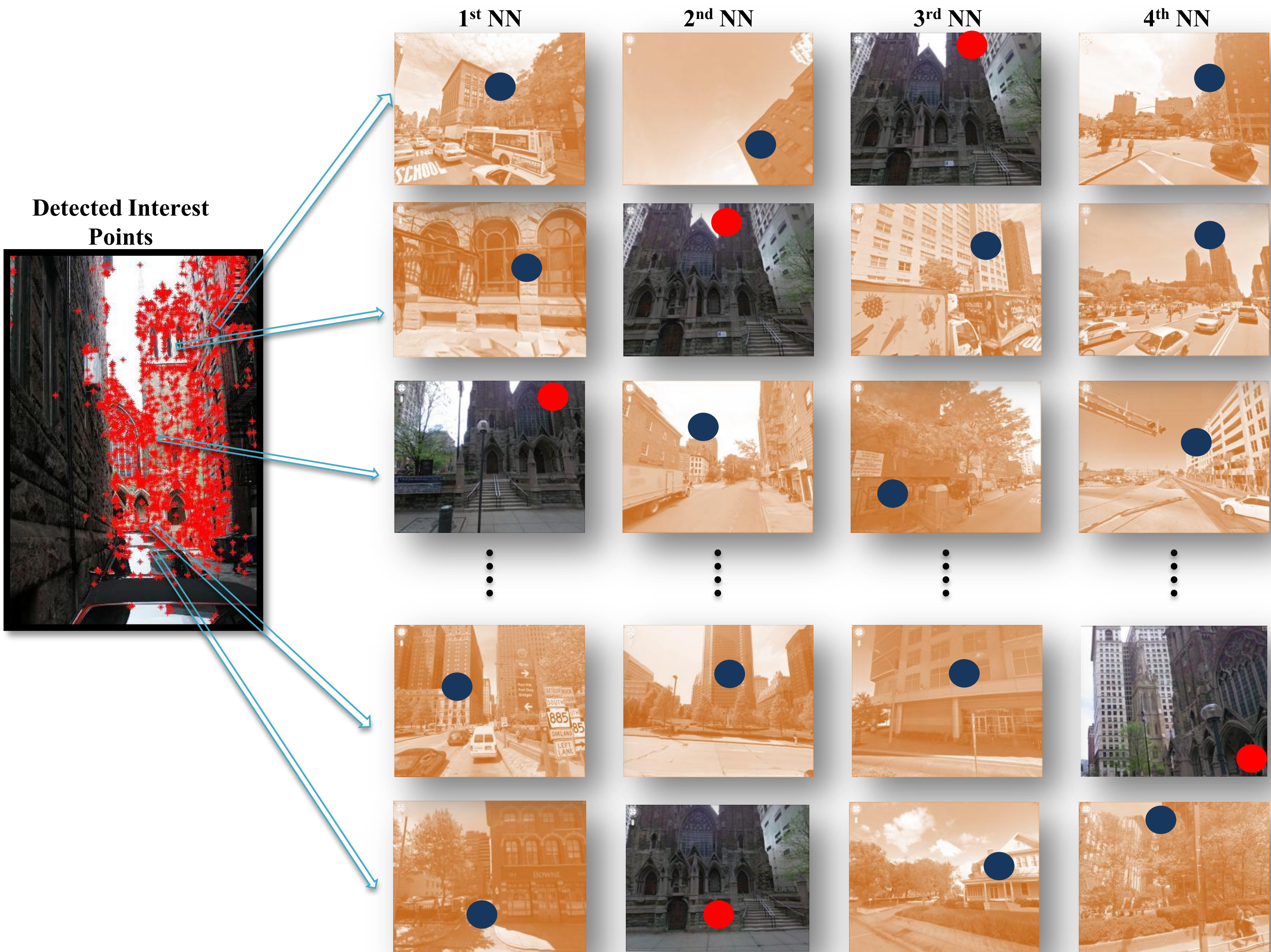
≠

- common in urban area
- Disambiguation using global features:
  - Global Color Histogram
  - GIST
  - Geo-tags

Using Multiple Nearest Neighbors

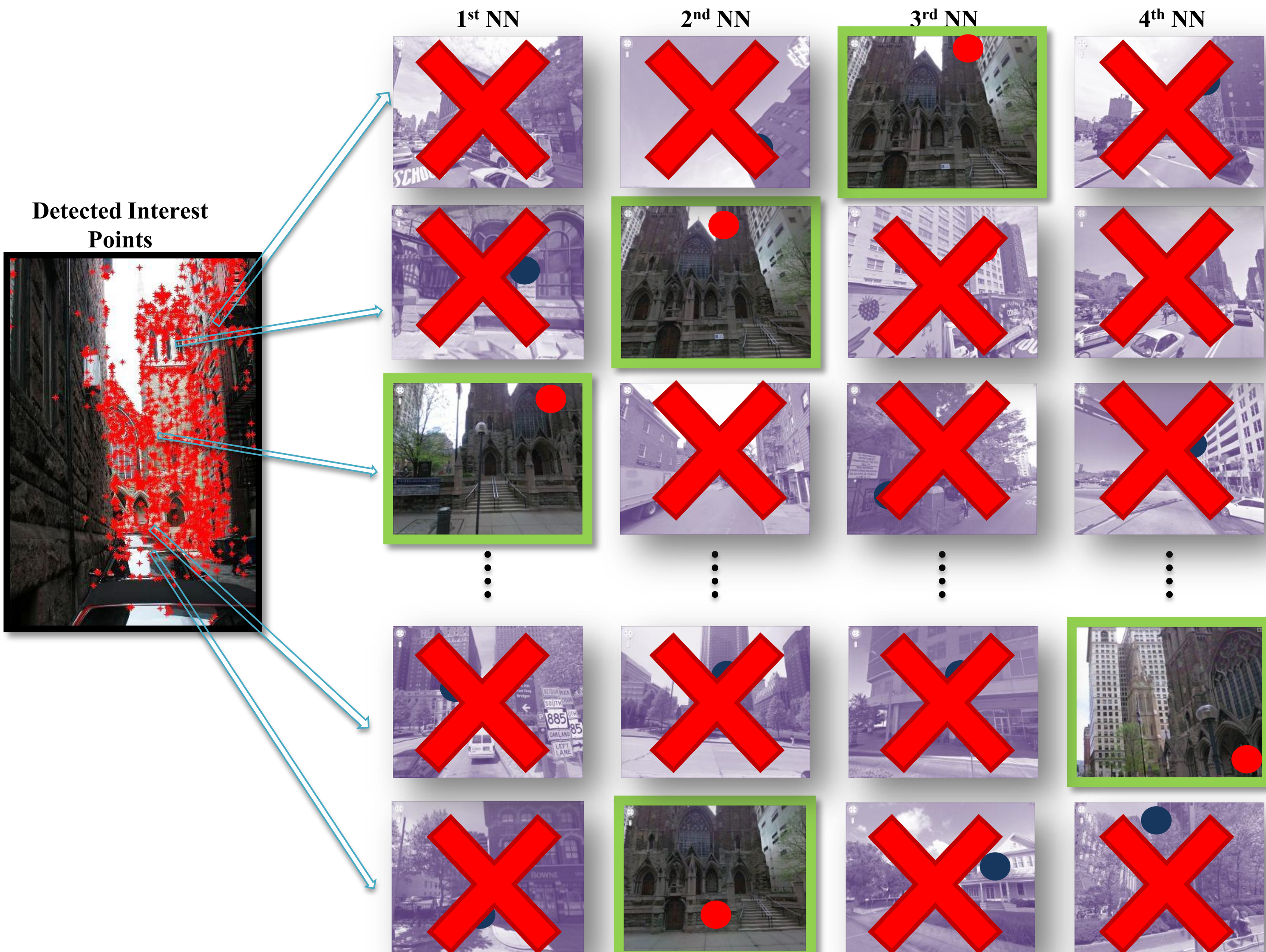


Using Multiple Nearest Neighbors



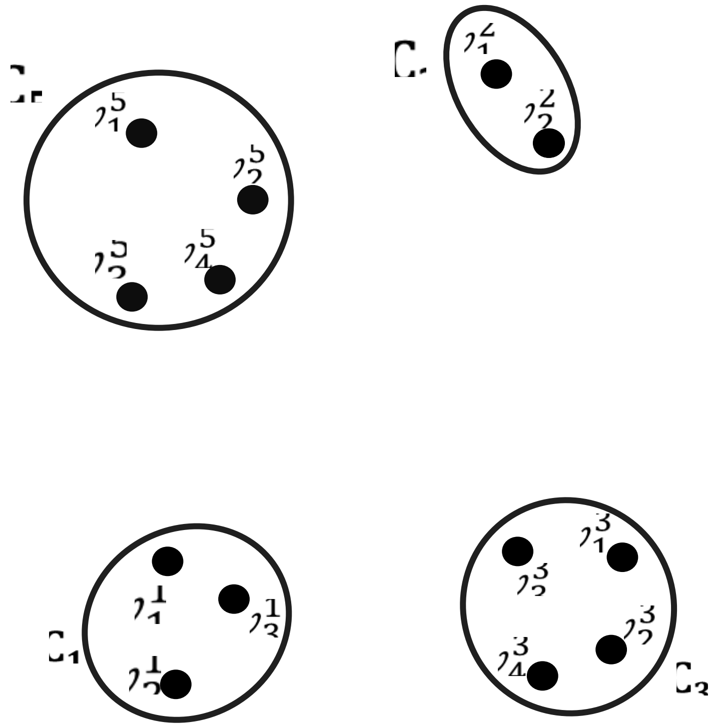
GMCP!

Using Multiple Nearest Neighbors



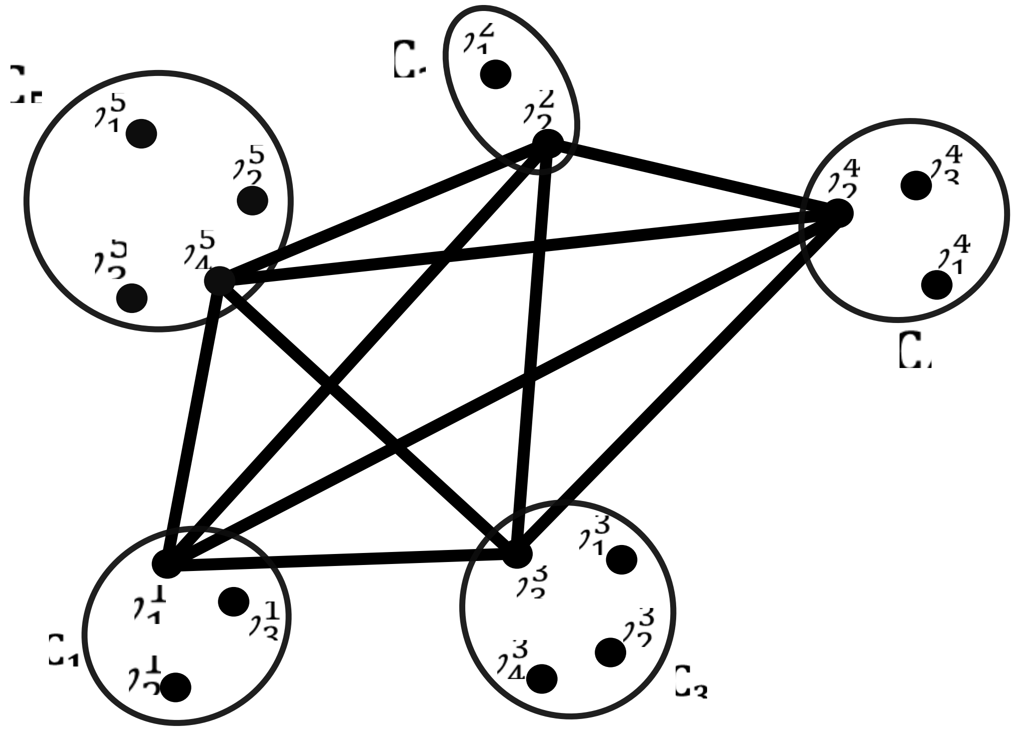
GMCP!

Generalized Minimum Clique Problem



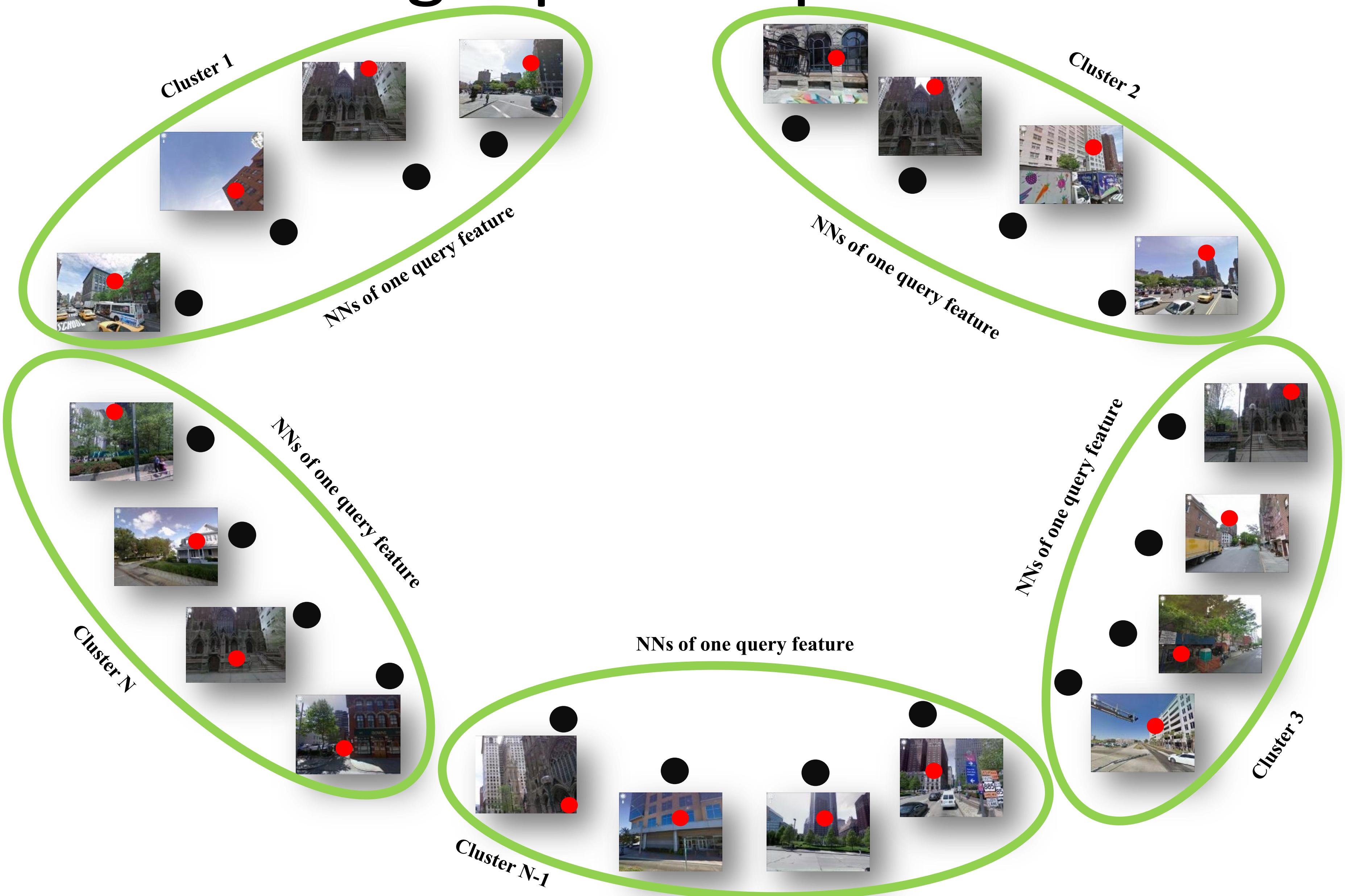
- Clusters of nodes.

Generalized Minimum Clique Problem

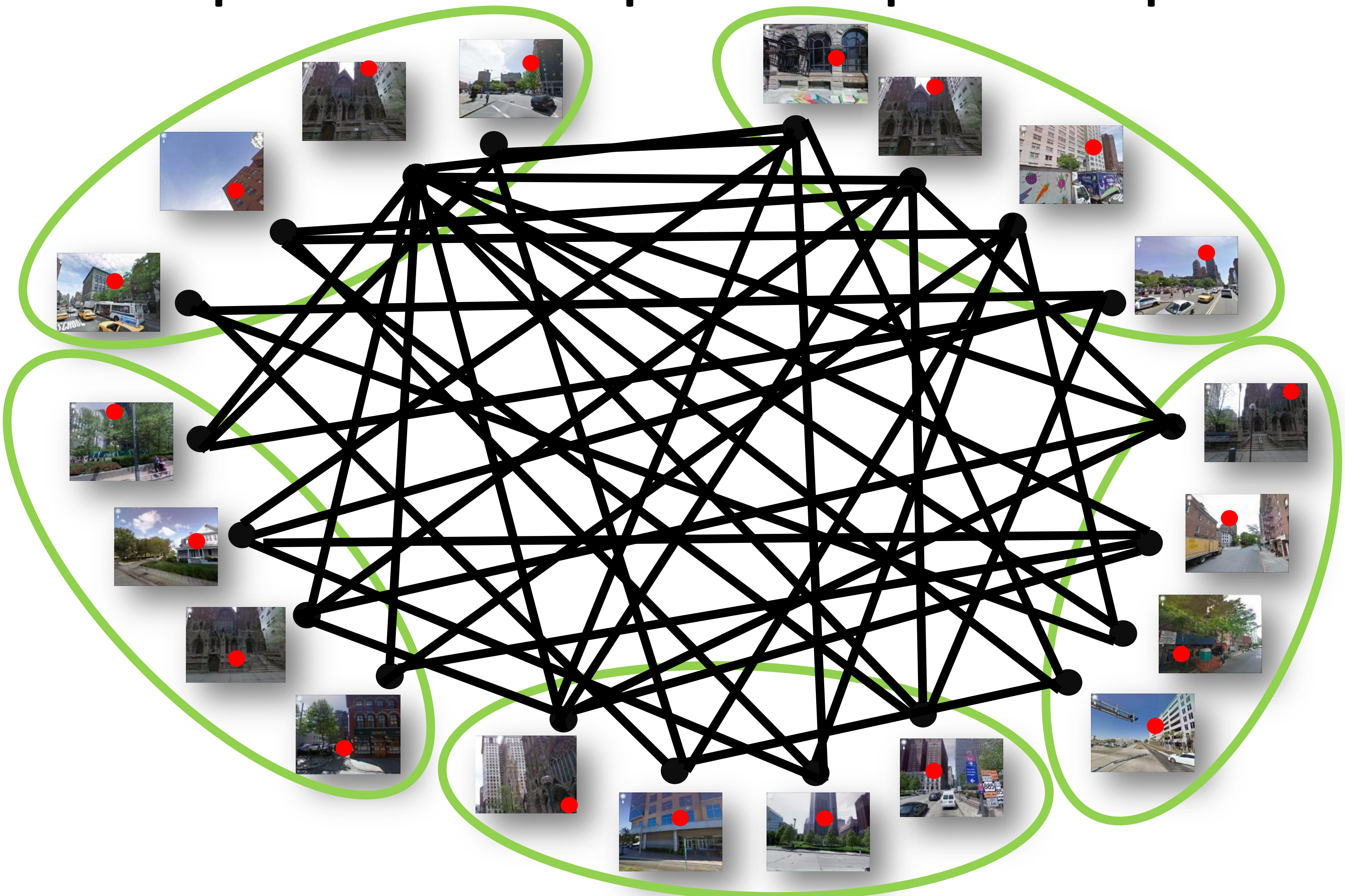


- Clusters of nodes.
- GMCP picks exactly one node out of each cluster
- The cost of the clique is minimized.

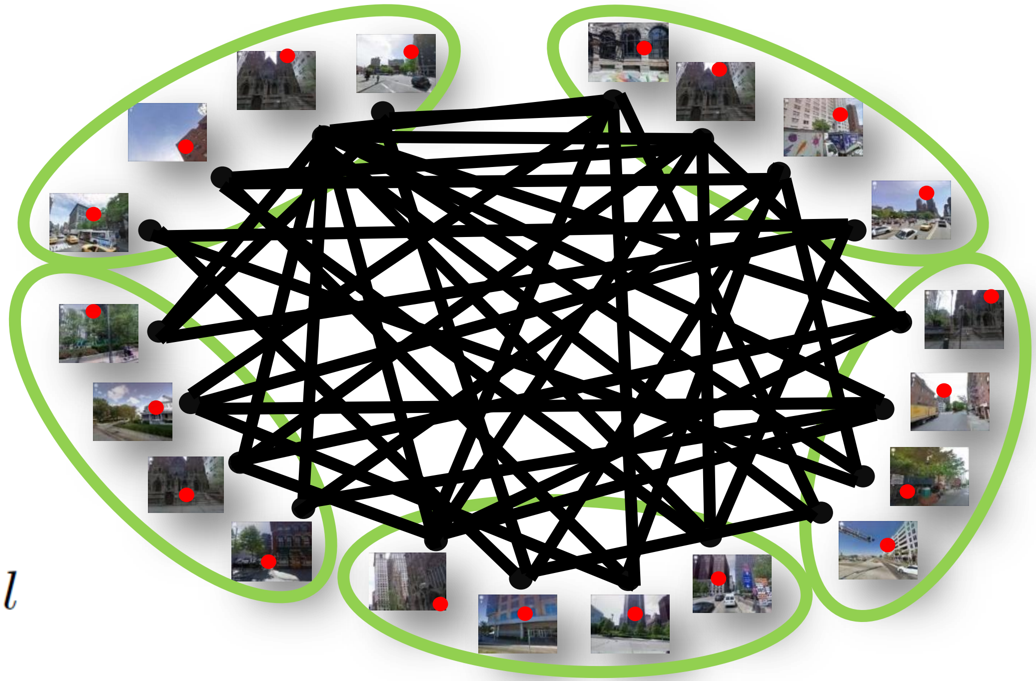
Forming Input Graph to GMCP



K-partite Complete Input Graph



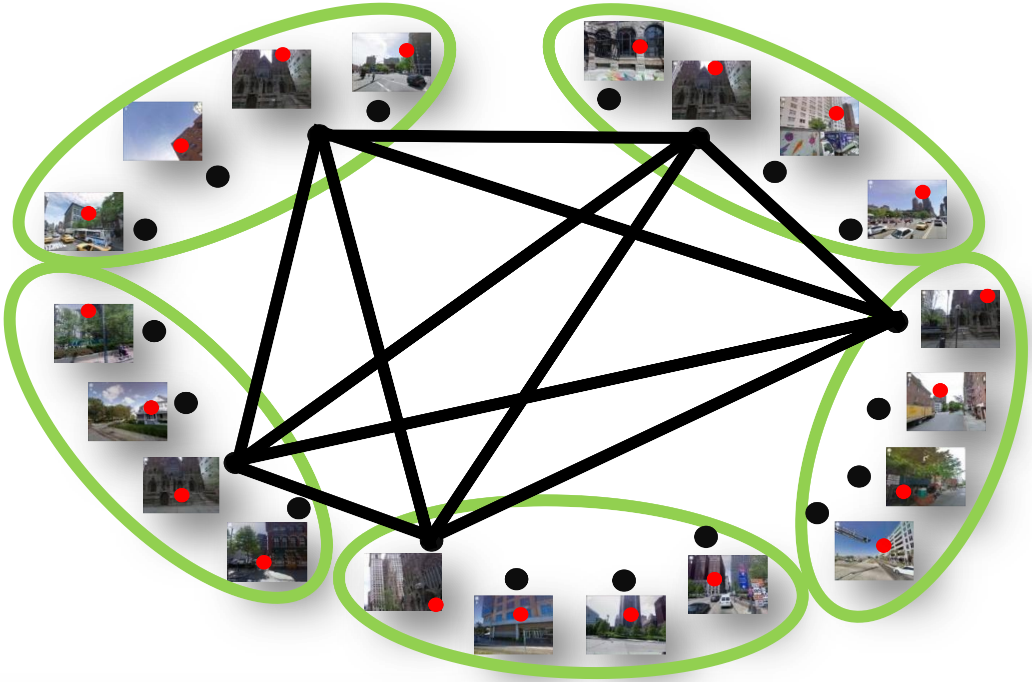
K-partite Complete Input Graph



$$\varpi(v_m^i) = \|q^i - \zeta(v_m^i)\|_l$$

$$w(v_m^i, v_n^j) = \|\rho(v_m^i) - \rho(v_n^j)\|_g$$

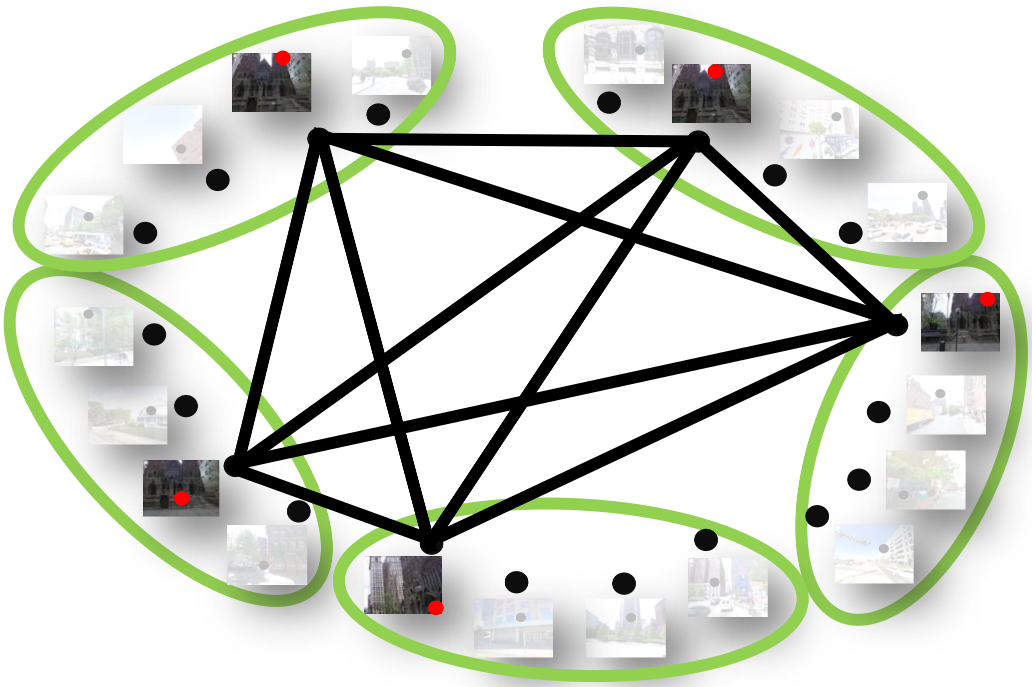
Generalized Minimum Clique



$$\hat{\mathbf{V}}_s = \underset{\mathbf{V}_s}{\arg\min} \; C(\mathbf{V}_s)$$

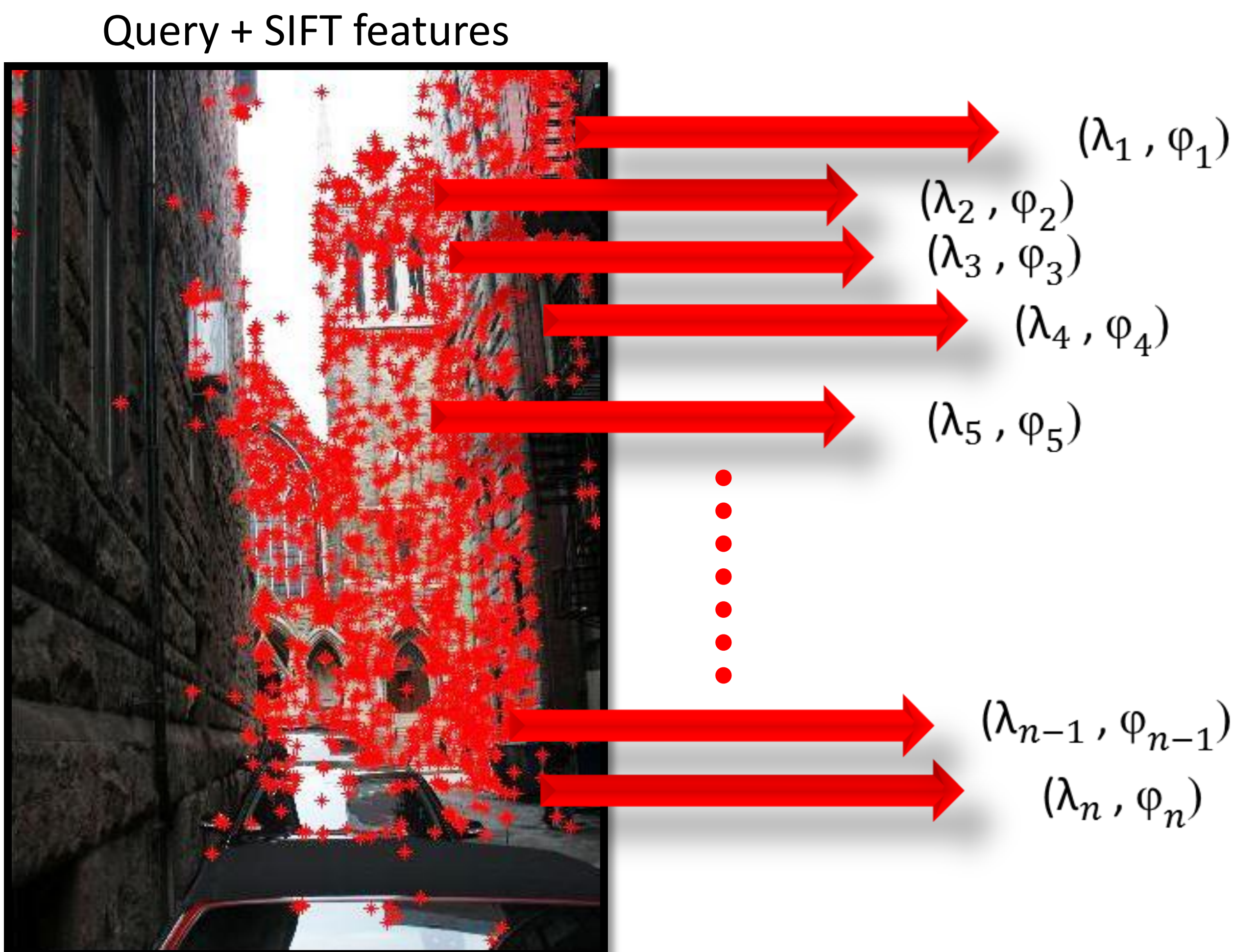
$$C(\mathbf{V}_s) = \frac{1}{2} \sum_{i=1}^{L} \sum_{\substack{j=1, \\ j \neq i}}^{L} \left(\frac{1}{2} \alpha \overbrace{\Big(\varpi(\mathbf{V}_s(i)) + \varpi(\mathbf{V}_s(j)) \Big)}^{\text{local features}} + (1 - \alpha) \underbrace{w(\mathbf{V}_s(i), \mathbf{V}_s(j))}_{\text{global features}} \right)$$

Generalized Minimum Clique

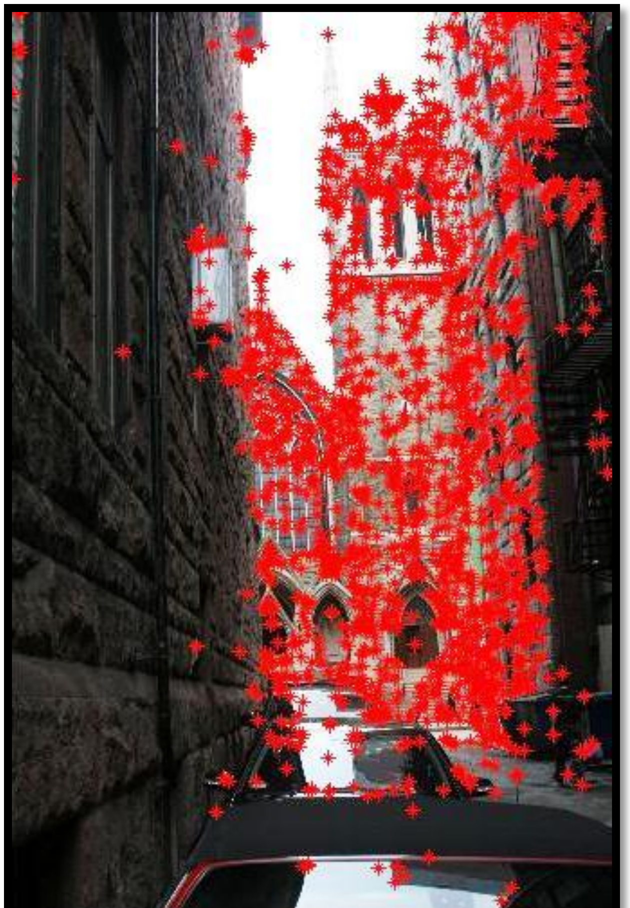


- Subset of NNs with maximum agreement in local and global features

From feature correspondences to GPS location

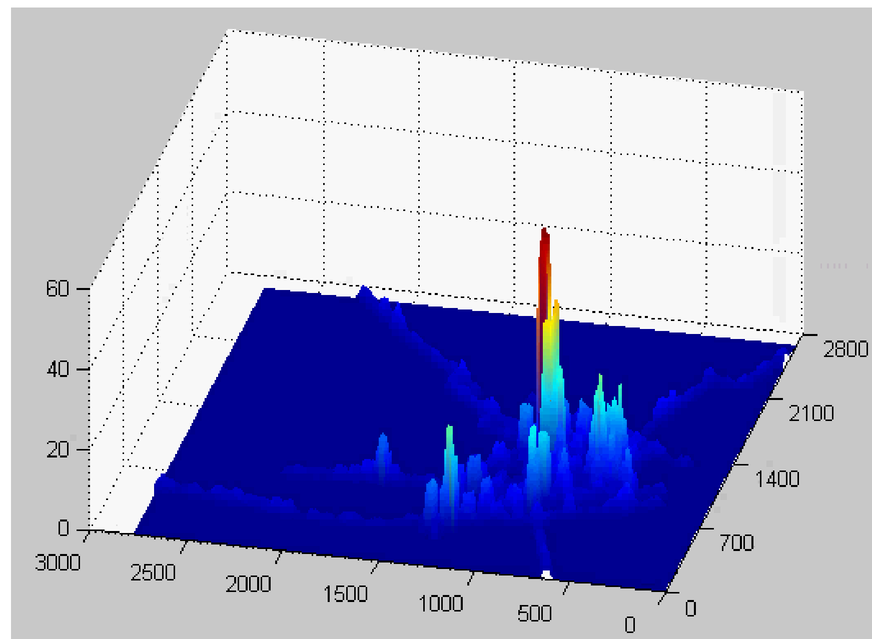


From feature correspondences to GPS location



Geo-location PDF

Vertical Axis: Number of Votes



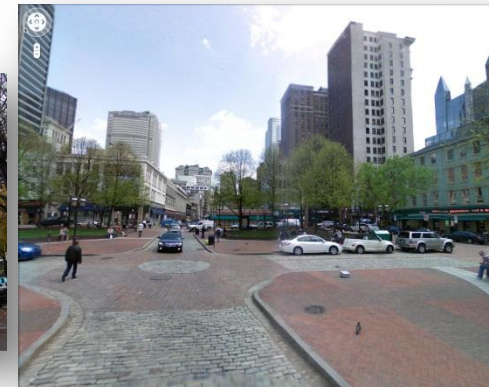
Horizontal Axes: Location Coordinates (x,y)

Geo-localization Results

Query

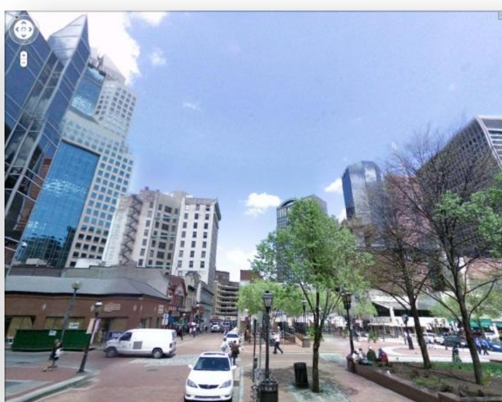
Match – Error: 7.6 m

Query



Match – Error: 6.9 m

Query



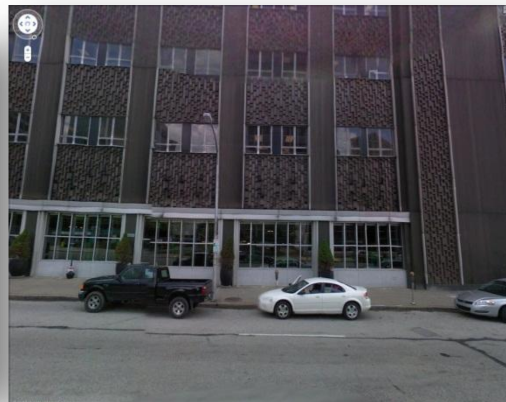
Match – Error: 308.1 m

Query

Match – Error: 59.3 m

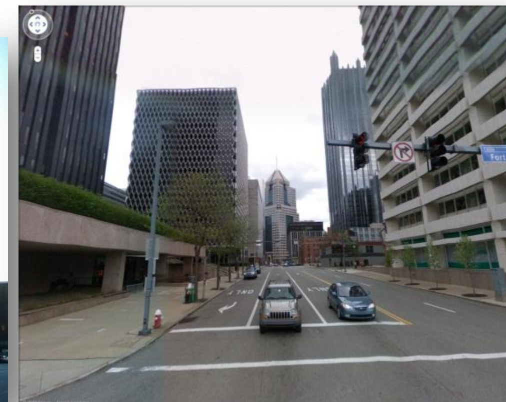
Geo-localization Results

Query



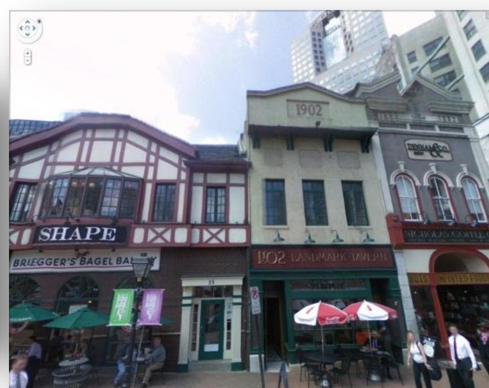
Match – Error: 64.2 m

Query



Match – Error: 159.8 m

Query

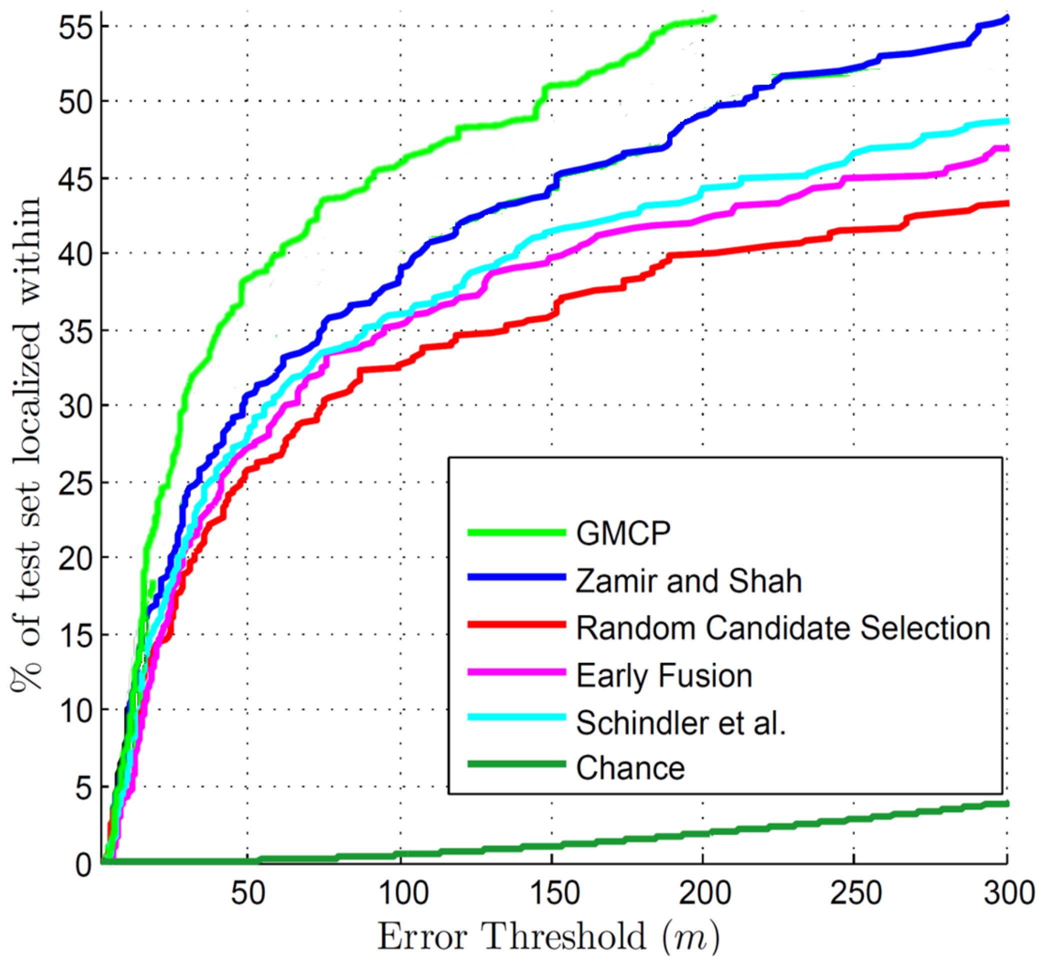


Match – Error: 6.4 m

Query

Match – Error: 199.8 m

Geo-localization Results



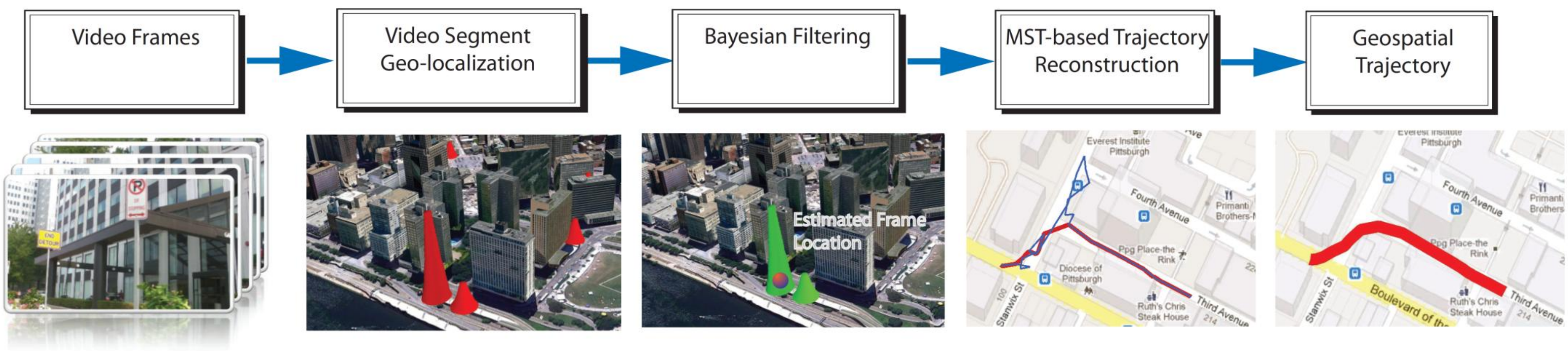
- **Automatic Geo-localization:**
  - Image geo-localization using Generalized Graphs
  - Video Geo-localization and trajectory extraction
- **Robust Refinement of geo-location using Random Walks**
- **Location-Aware image understating:**
  - Location-aware object detection
  - Precise recognition of storefronts in images

Paper:

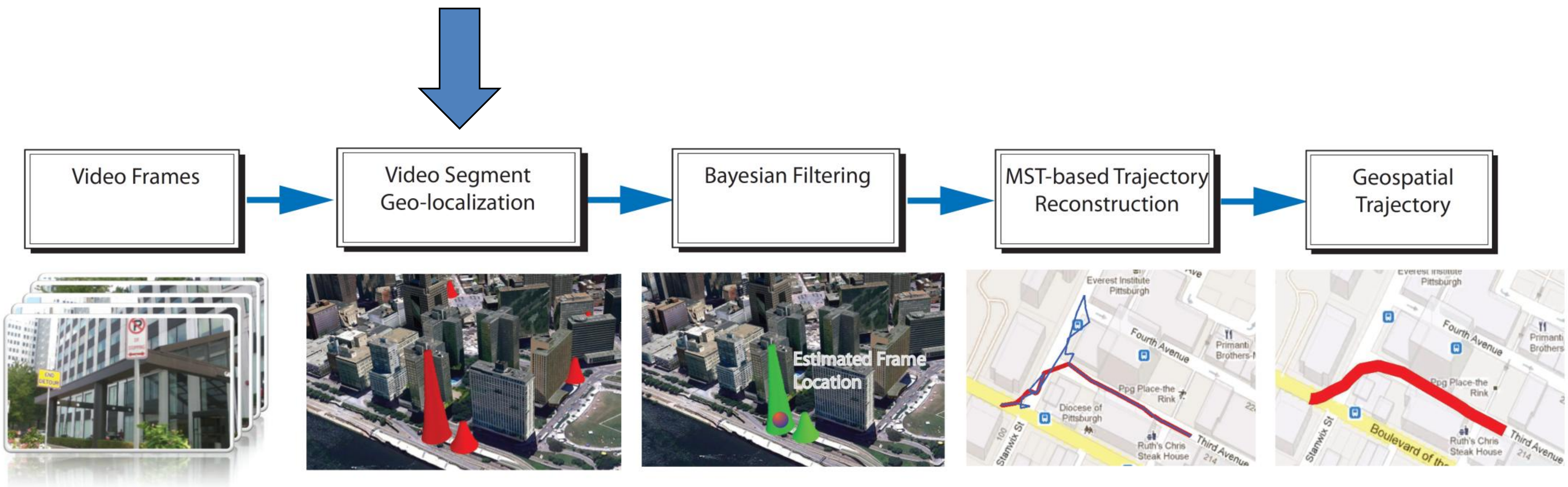
City Scale Geo-spatial Trajectory Estimation of a Moving Camera, In ***CVPR***, 2012.

How about Videos?!

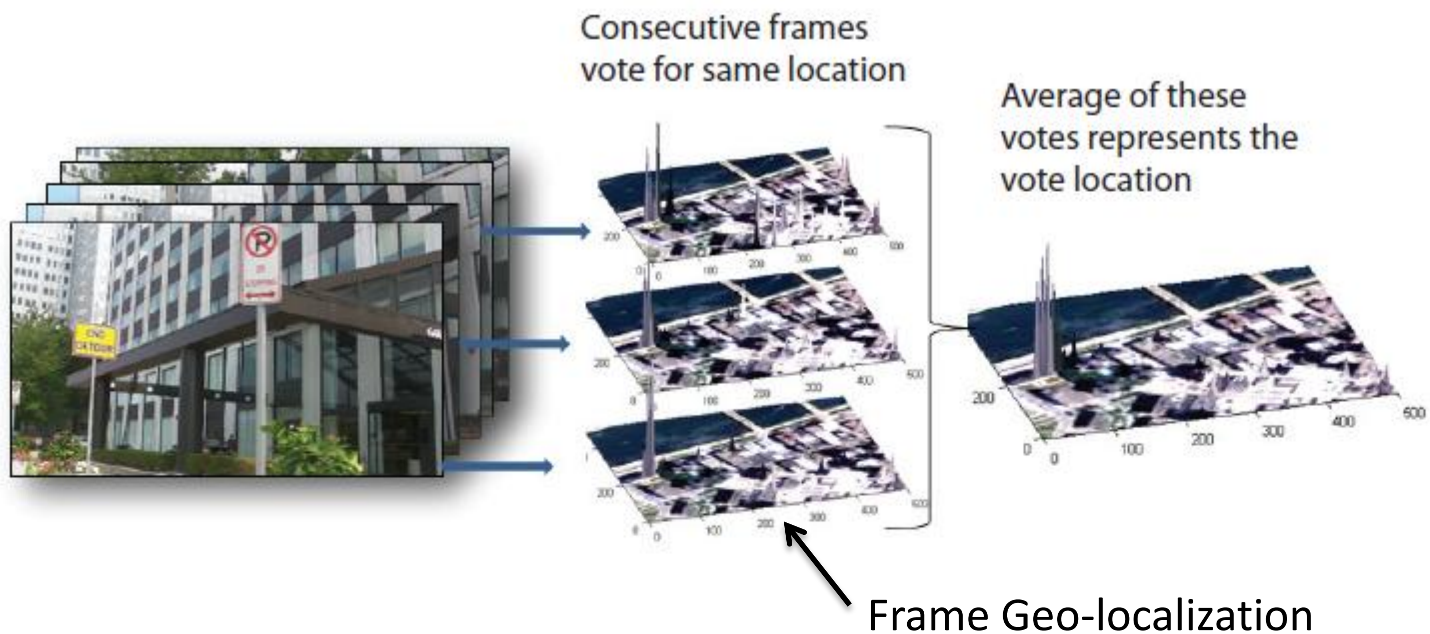
Our Approach



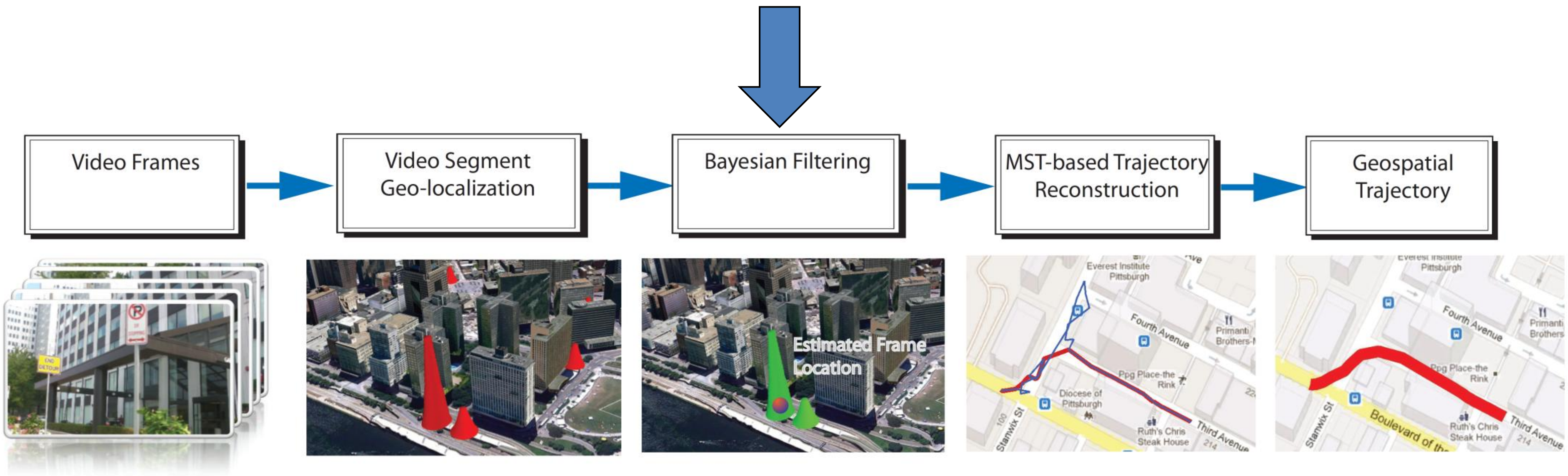
Segment Geo-Localization



Segment Geo-Localization



Bayesian Filtering



Bayesian Filtering

- Enforces temporal consistency.

Bayesian Filtering

- Enforces temporal consistency.

$$p(x_t|Z_t) = \frac{\boxed{p(z_t\,|x_t)}\; p(x_t\,|z_{t-1})}{c}$$

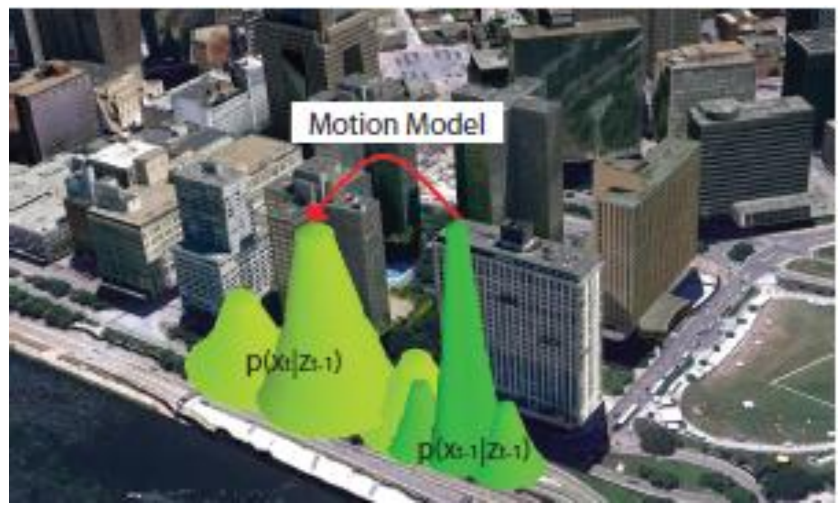
- State (unobserved) x = [lat , long]
- Measurement (observed) z

Likelihood
(Current Segment)

Bayesian Filtering

$$p(x_t|Z_t) = \frac{p(z_t\,|x_t)\;\; p(x_t\,|z_{t-1})}{c}$$

Prediction of the state from the previous state

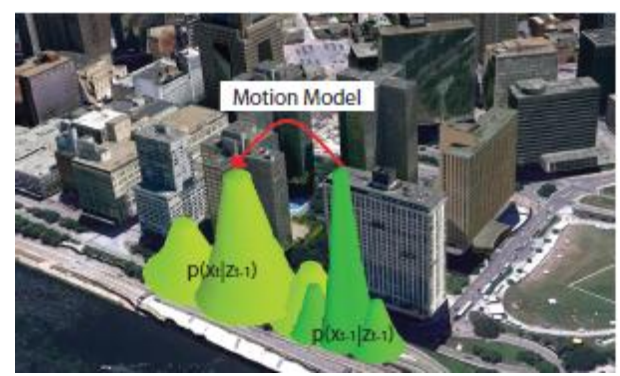


Prediction

Bayesian Filtering

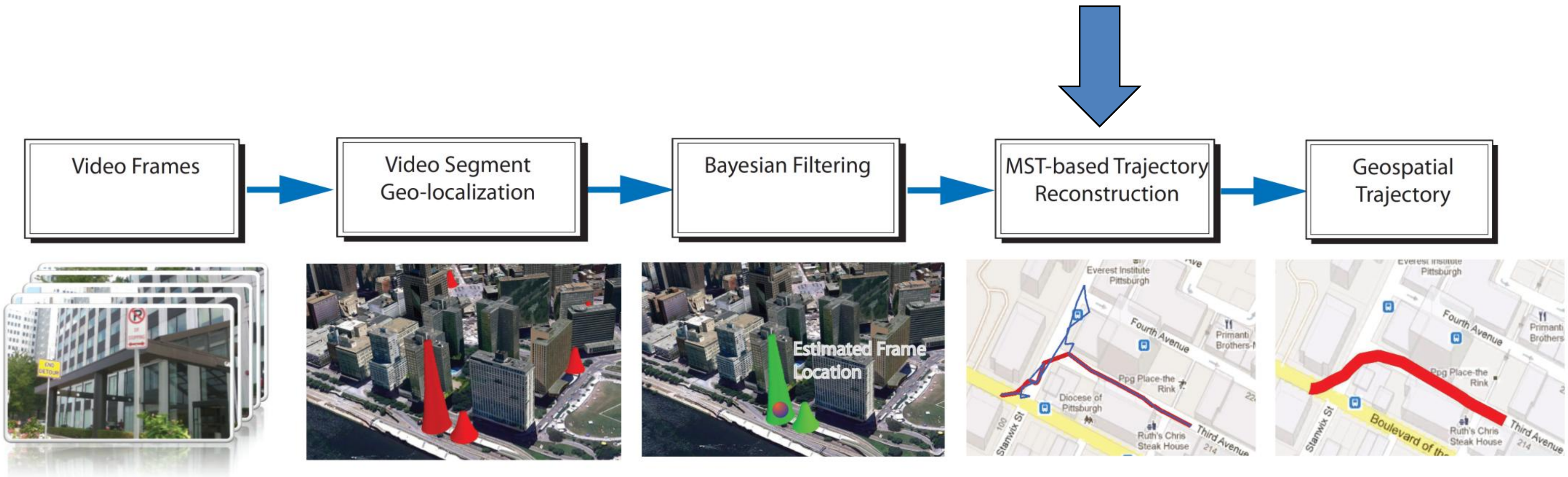
$$p(x_t|Z_t) = \frac{p(z_t\,|x_t)\;\;p(x_t\,|z_{t-1})}{c}$$

Likelihood
(Current Segment)



Prediction
(previous Segment)

MST-based Trajectory Reconstruction



Geo-localization of a YouTube Video

- **Automatic Geo-localization:**
  - Image geo-localization using Generalized Graphs
  - Video Geo-localization and trajectory extraction
- **Robust Refinement of geo-location using Random Walks**
- **Location-Aware image understating:**
  - Location-aware object detection
  - Precise recognition of storefronts in images

Paper:

GPS-Tag Refinement using Random Walks with an Adaptive Damping Factor. In ***CVPR***, 2014.

Why GPS-tag Refinement?

- What if the image is already geo-tagged?
  - Internal GPS, WPS, Cell Signal Positioning, Manual tagging.
  - **Known issue:** error in user shared geo-tags
    - Mean=428 meters in 20% of data.

Block Diagram

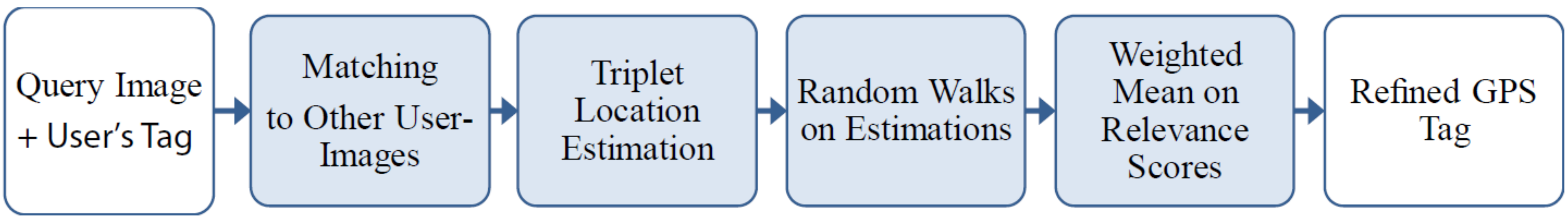
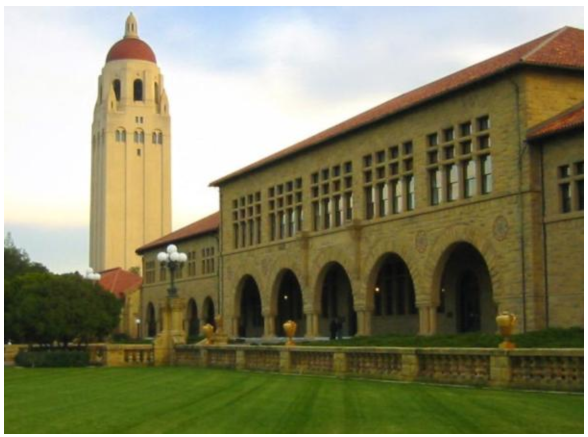


Image Matching

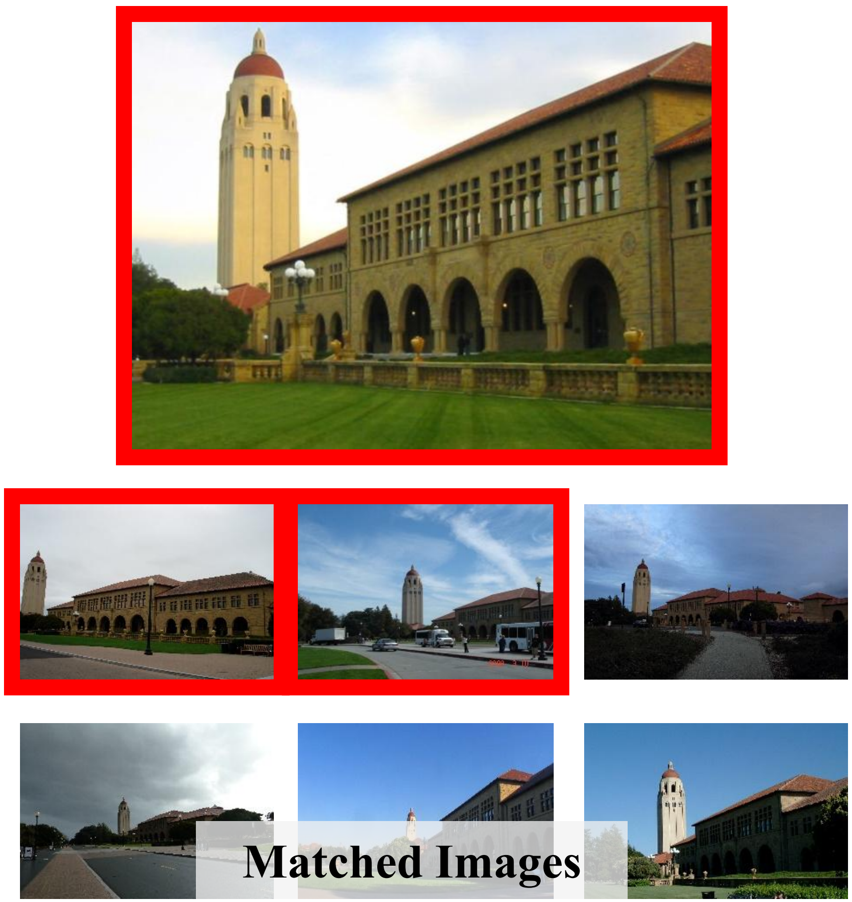
Query Image



Matched Images

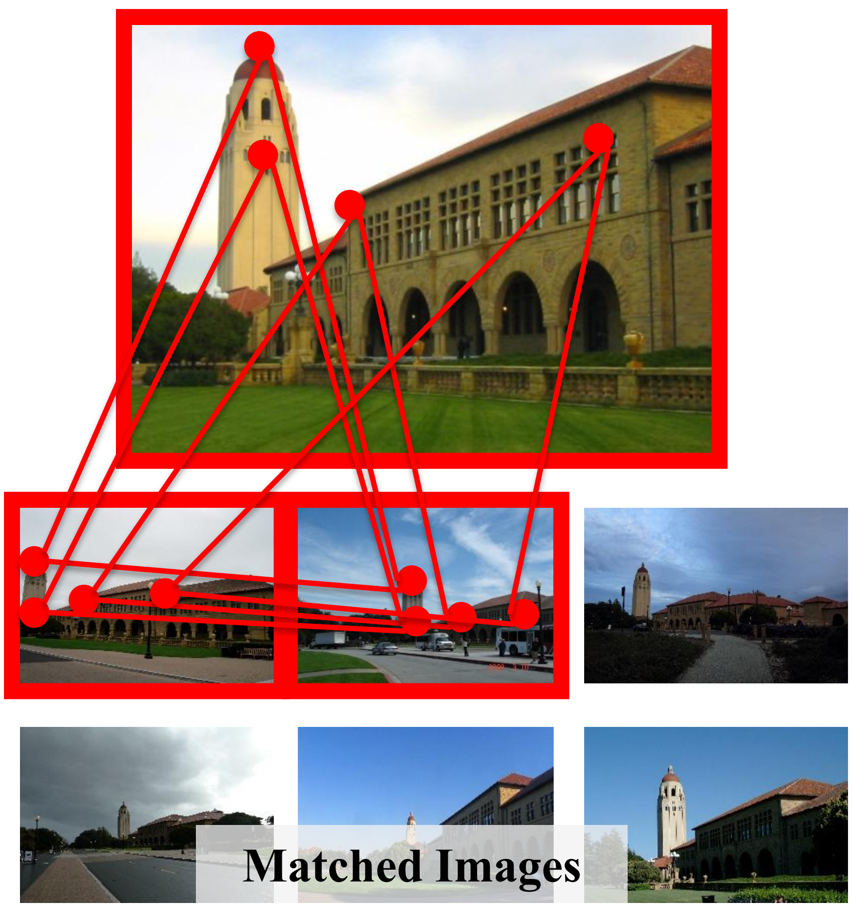
Triplet Location Estimation

Query Image

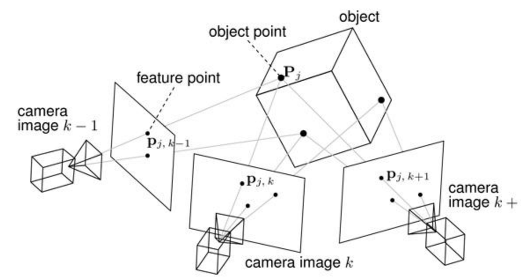


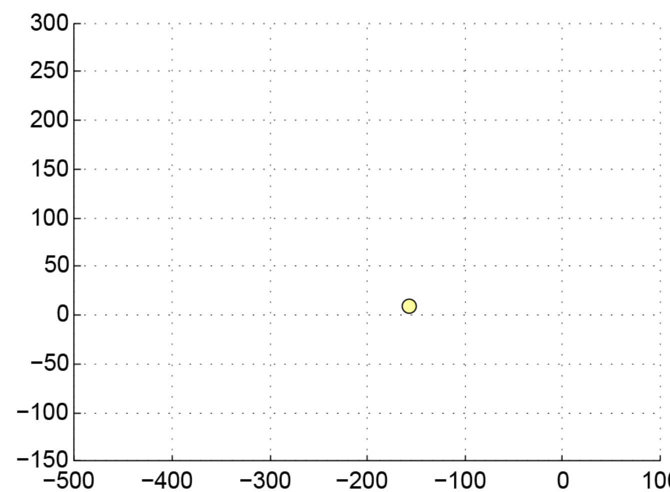
Triplet Location Estimation

Query Image



Trifocal Tensor

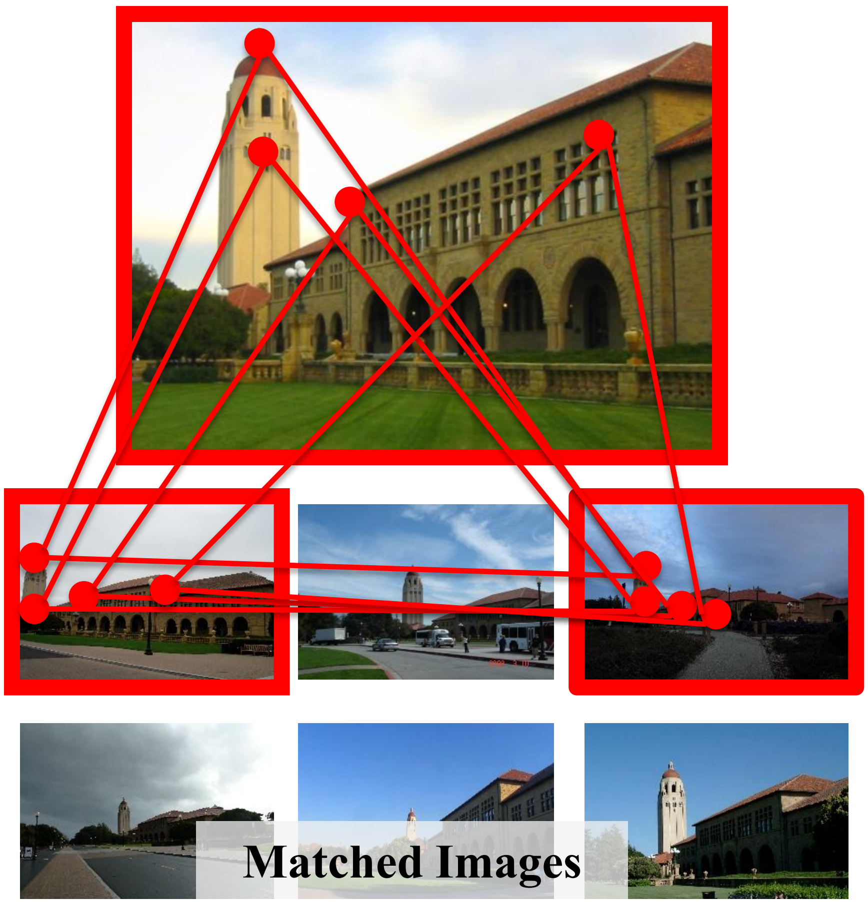




Estimations on Map (ENU Metric System)

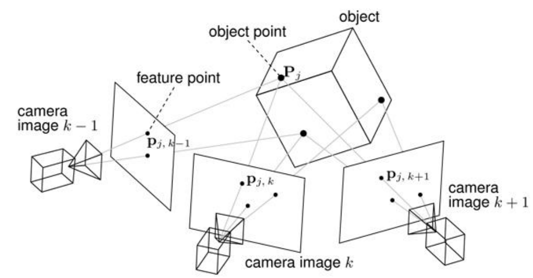
Triplet Location Estimation

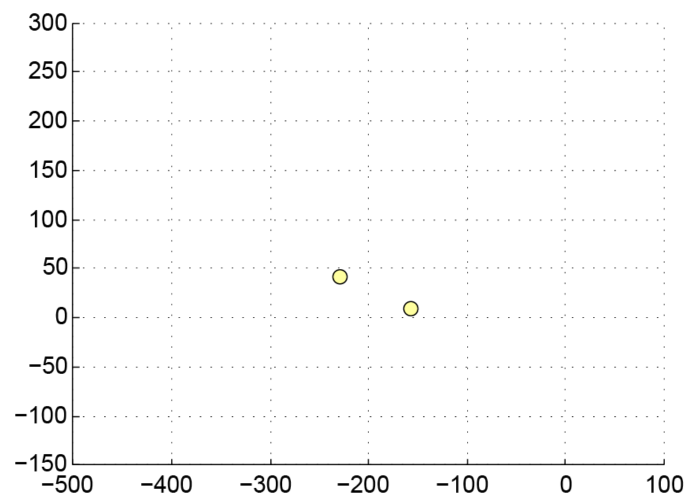
Query Image



Matched Images

Trifocal Tensor

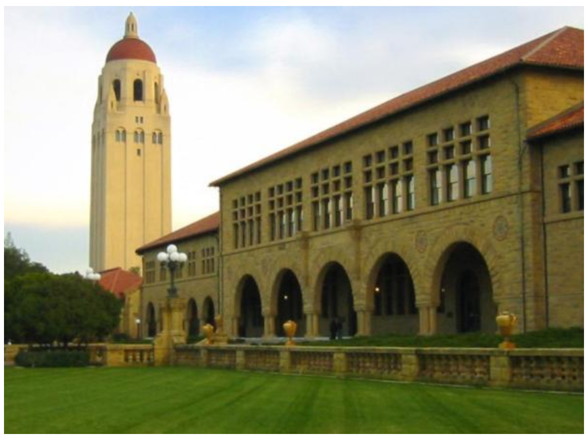




Estimations on Map (ENU Metric System)

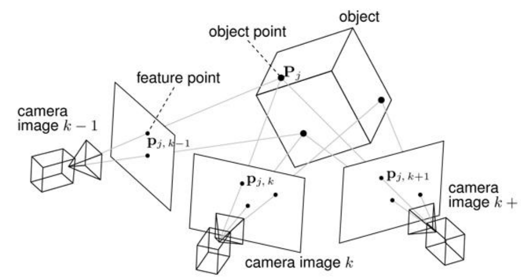
Triplet Location Estimation

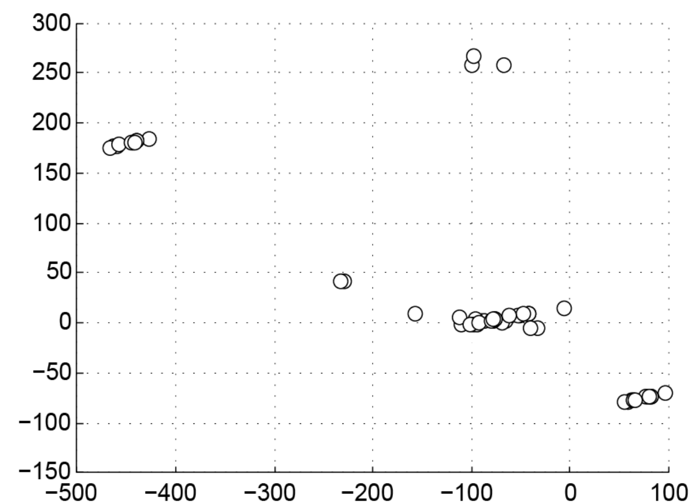
Query Image



Matched Images

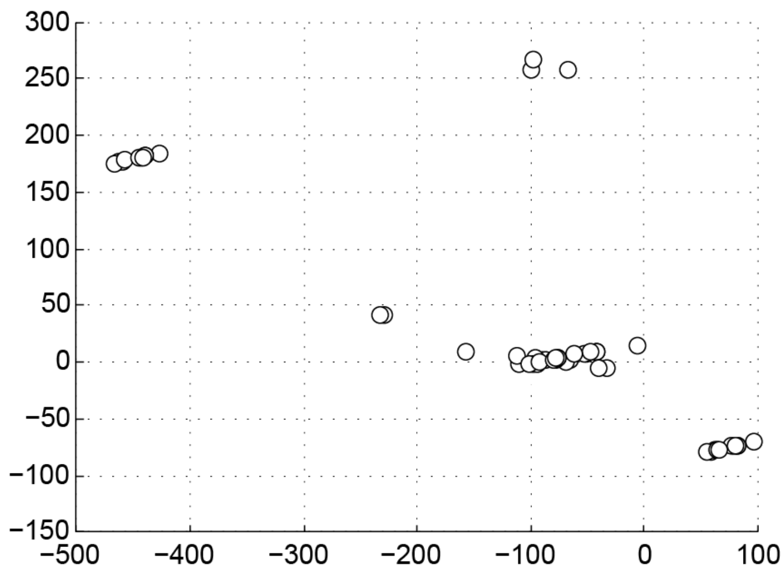
Trifocal Tensor

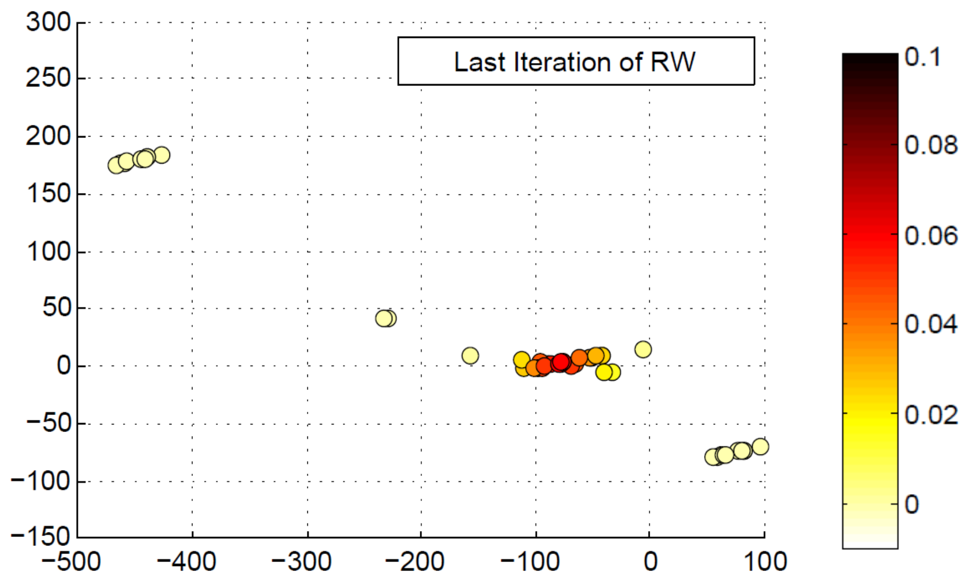




Estimations on Map (ENU Metric System)

Random Walks on GPS estimations





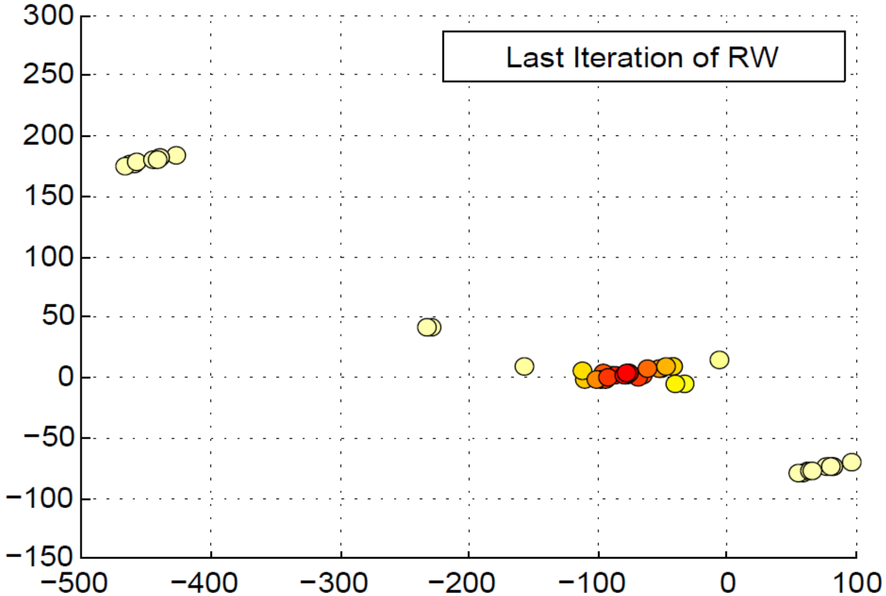
$$p(i,j) = \frac{e^{-\sigma \|\boldsymbol{g}_i - \boldsymbol{g}_j\|_2}}{\sum_{k=1}^{\lambda} e^{-\sigma \|\boldsymbol{g}_i - \boldsymbol{g}_k\|_2}}$$

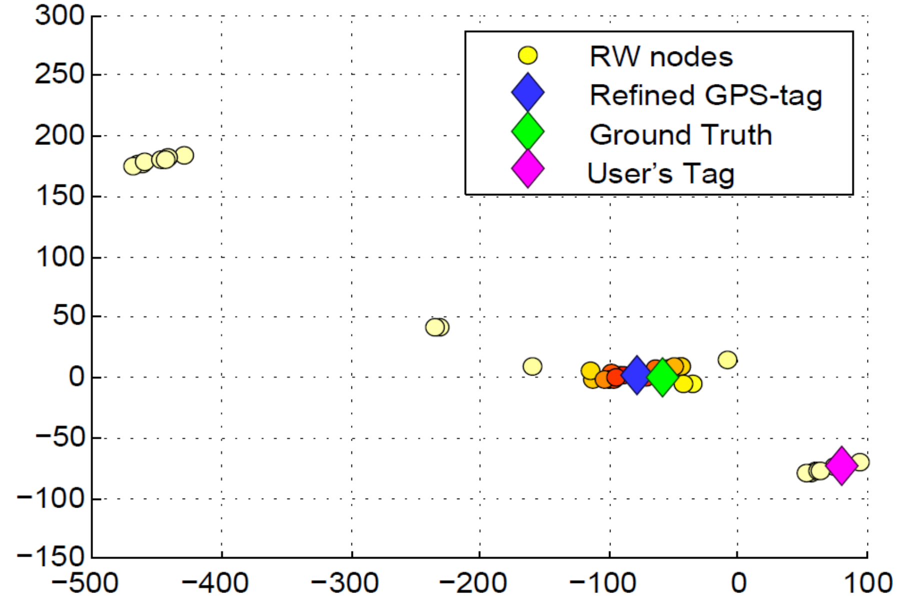
$$x_{(k+1)}(j) = \sum_{i=1}^{\lambda} \alpha \;\; x_k(i)p(i,j) + (1-\alpha)v(j)$$

Refined Location Estimation

- Weighted mean using Stationary relevance scores:

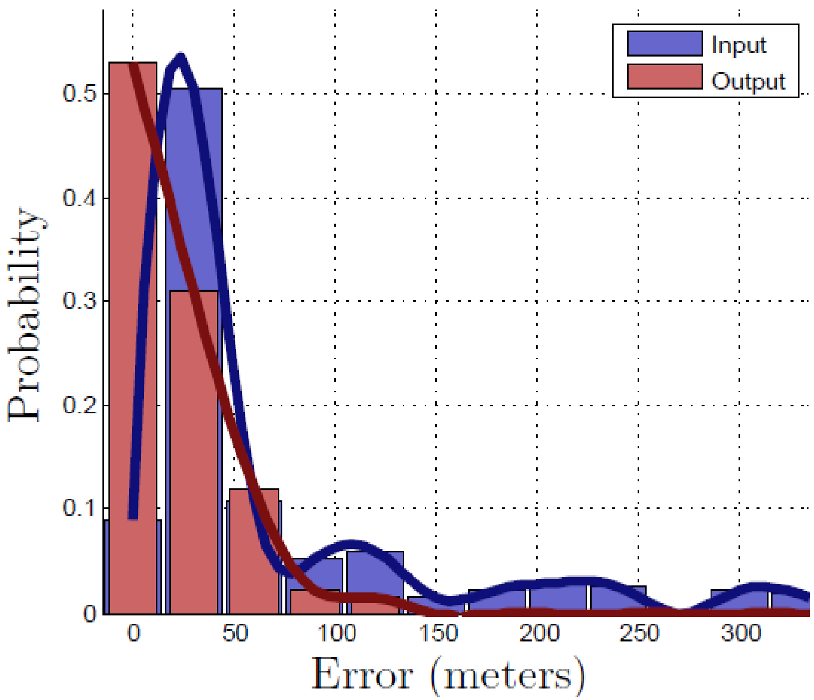
$$\hat{g} = \sum_{i=1}^{\lambda} g_i x_\pi(i)$$

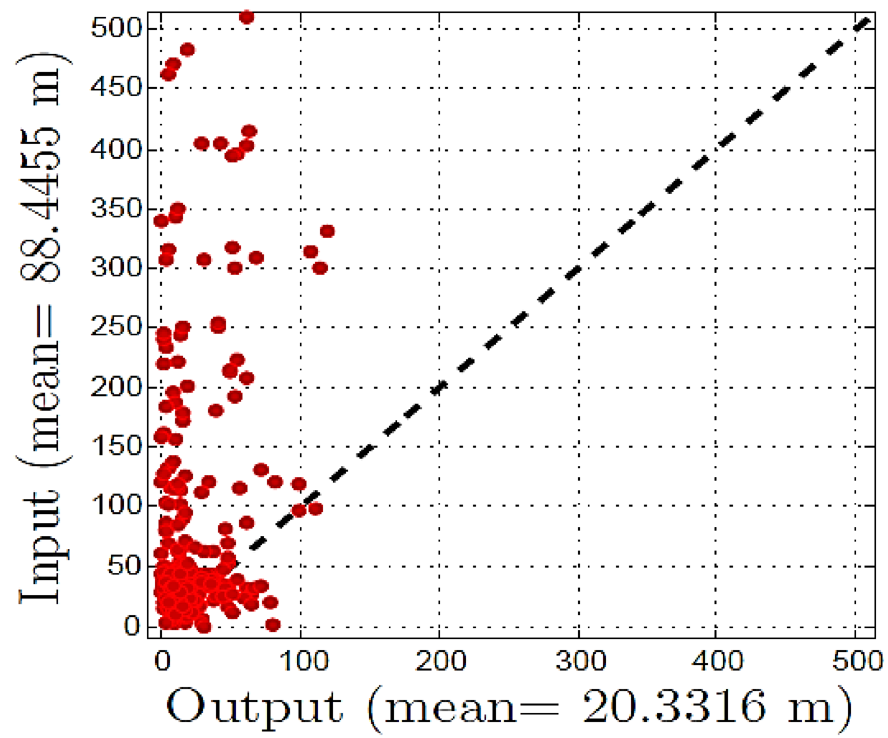




Experimental Results

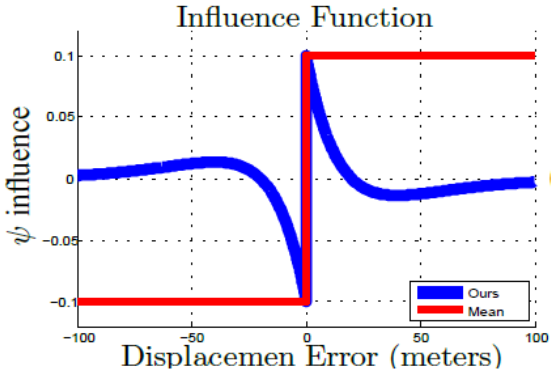
- 18075 User shared Images. (Flicker, Panoramio, Picasa)
- From San Francisco, CA; Pittsburgh, PA and Washington, DC.
- Test Set: 500 randomly selected subset.

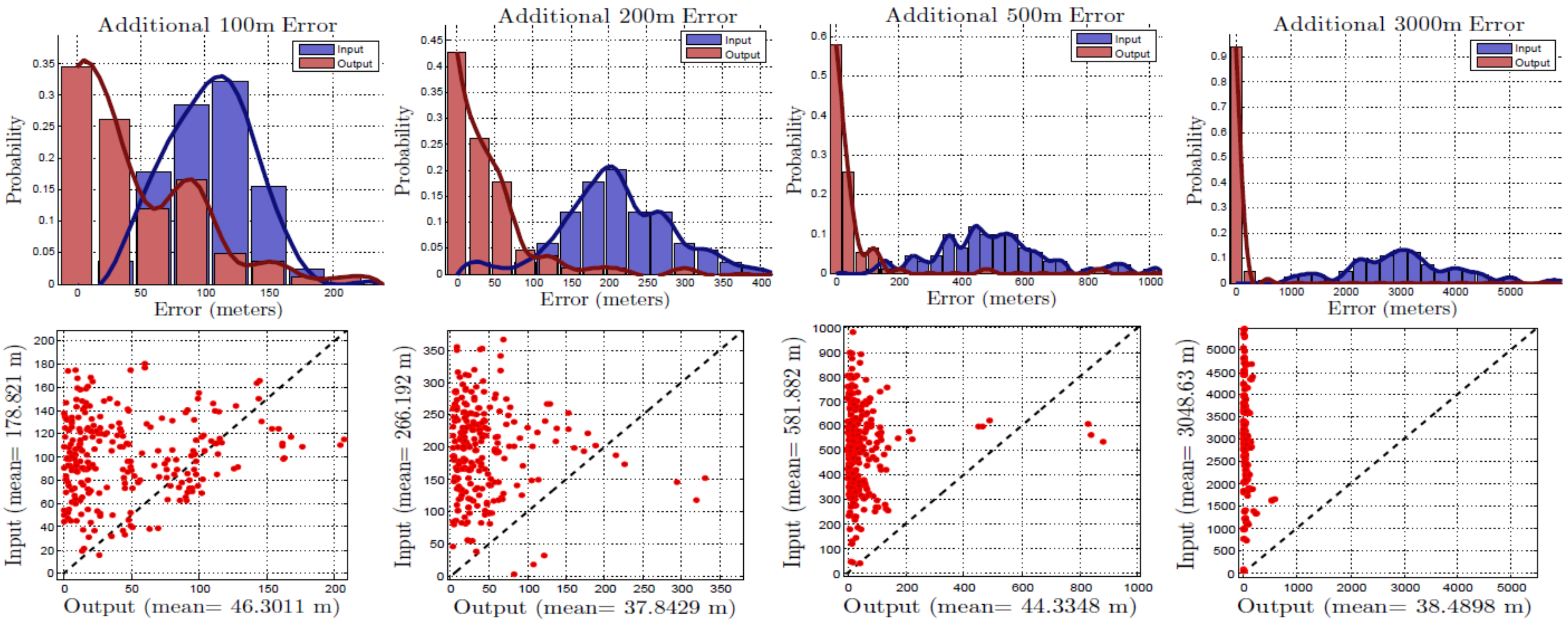




Robustification Test

- Larger amount of noise in the input:
- Adaptive Damping Factor was used.





- **Automatic Geo-localization:**
    - Image geo-localization using Generalized Graphs
    - Video Geo-localization and trajectory extraction
- **Robust Refinement of geo-location using Random Walks**
- **Location-Aware image understating:**
    - Location-aware object detection
    - Precise recognition of storefronts in images

Paper:

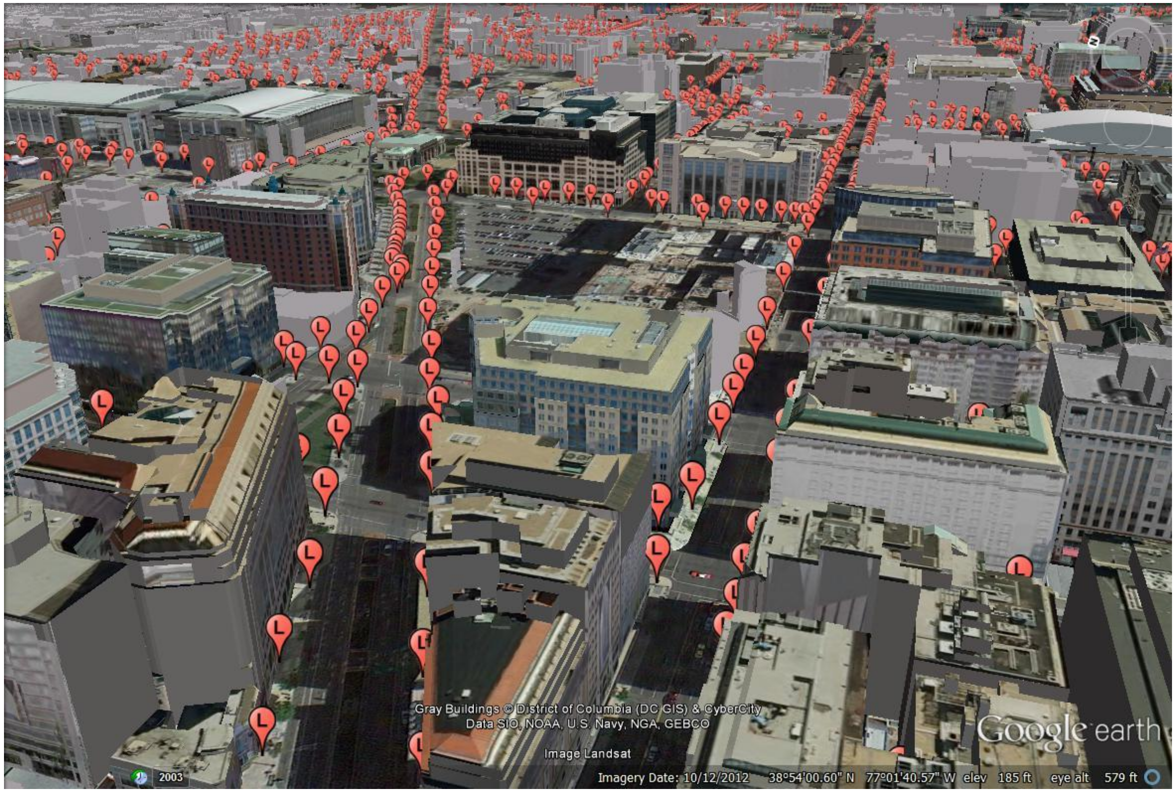
GIS-Assisted Object Detection and Geospatial Localization. In ***ECCV***, 2014.

Location-aware Image Understanding

- Geo-localization:
  - Estimate the location of the image/video.
- Geo-tag is known?
  - *Location-aware Image Understanding.*
  - Most images will be geo-tagged in the future.
  - Particularly important for real world applications.

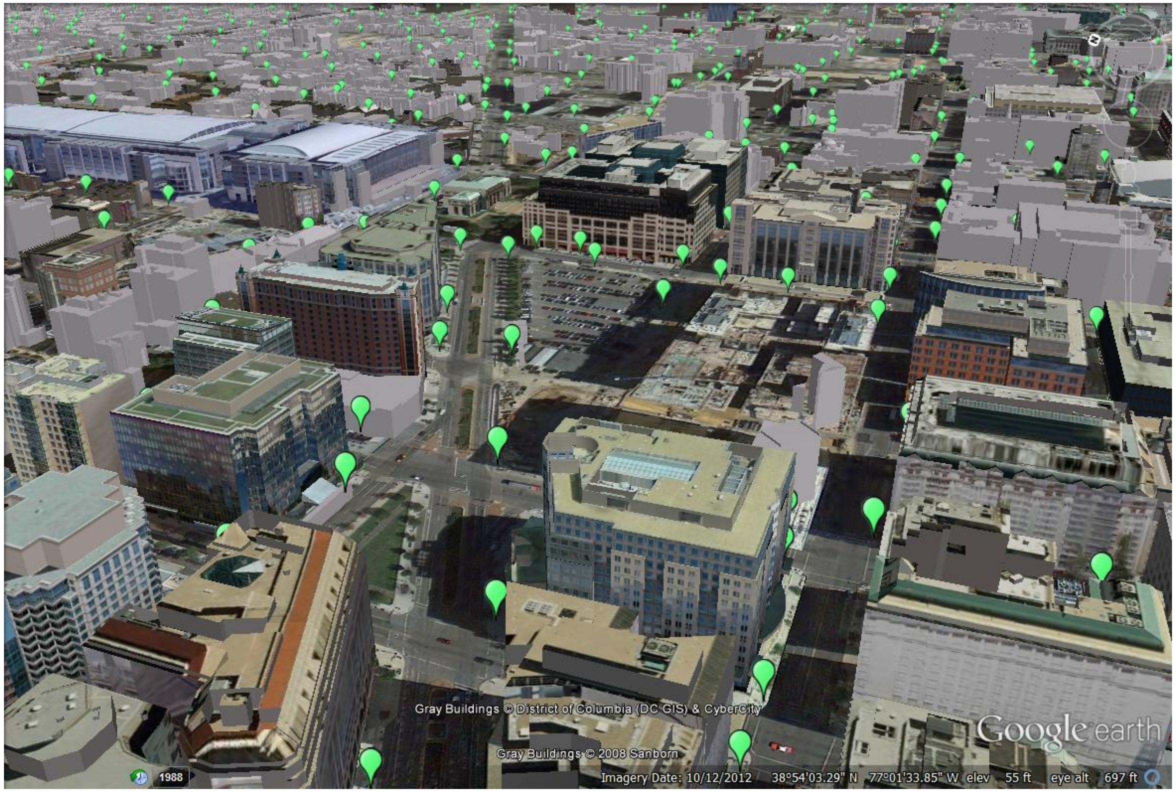
GIS Dataset

- e.g. Washington D.C.
  - Lamp posts



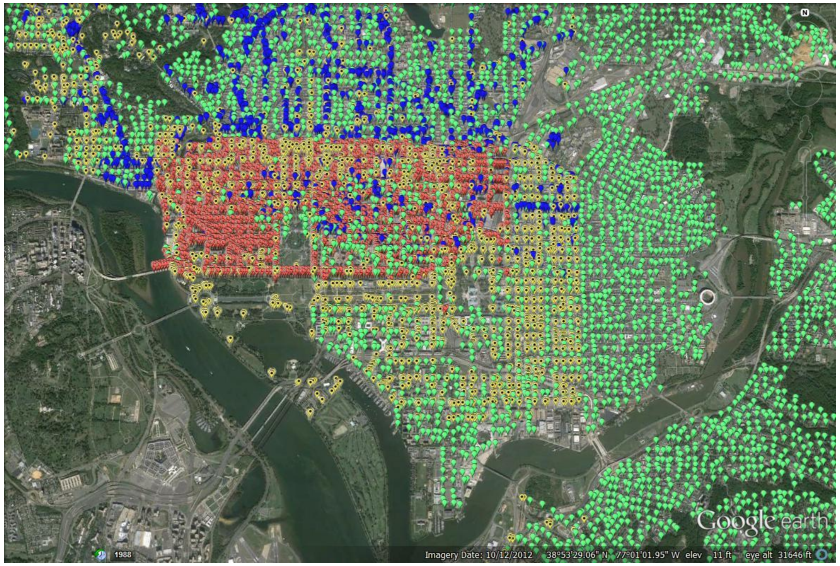
GIS Dataset

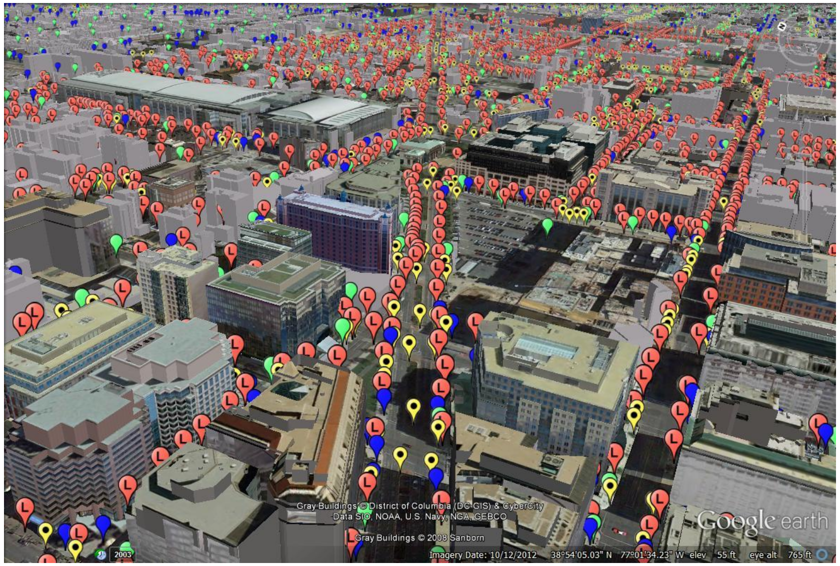
- e.g. Washington D.C.
  - Lamp posts
  - Fire Hydrants



GIS Dataset

- Locations of most stationary objects are documented!
- e.g. Washington D.C.
  - Buildings, Foliage, Road signs, ATMs, Fire Hydrants, Lamp posts, Cell/FM towers, Traffic Lights, Bus/subway stations, Trash cans.





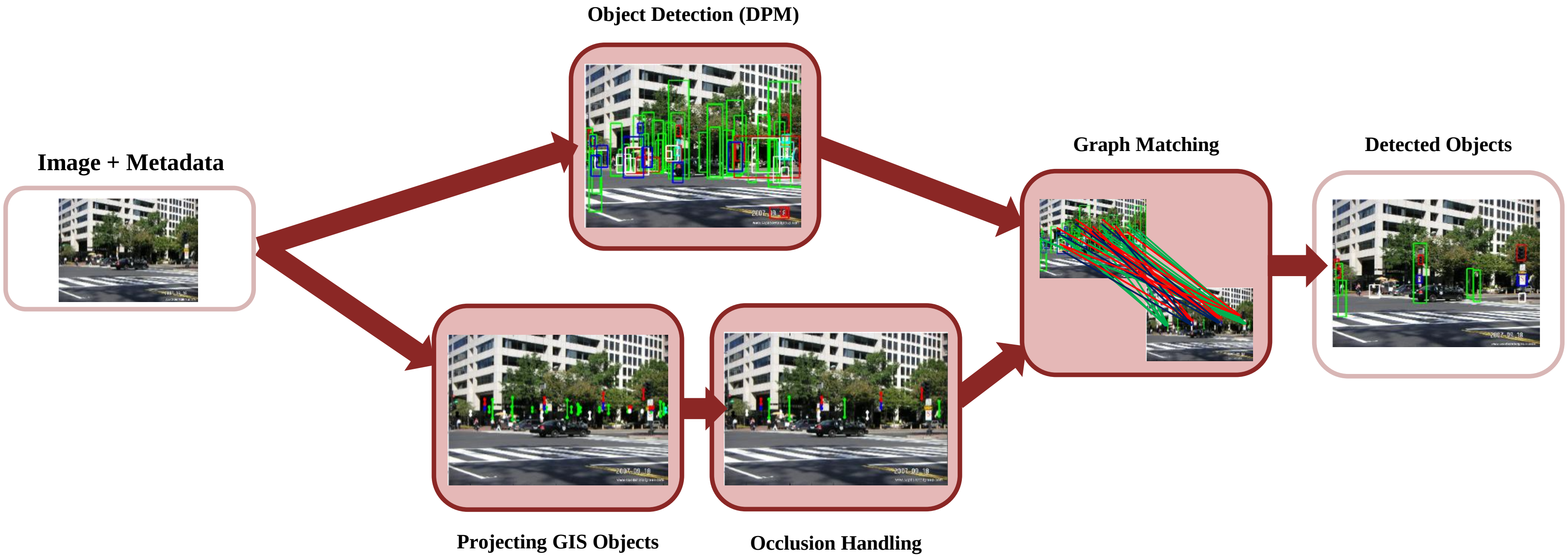
Fusion of Image content and GIS

Object Detectors

GIS

Lamp Post Road Sign Traffic Light

Location-aware Object Detection



Obtaining Priors from GIS

Camera View

All Projections

Non-occluded Projections

Camera Matrix

$$\boldsymbol{P} = \boldsymbol{C}[\boldsymbol{R} \mid \boldsymbol{T}]$$

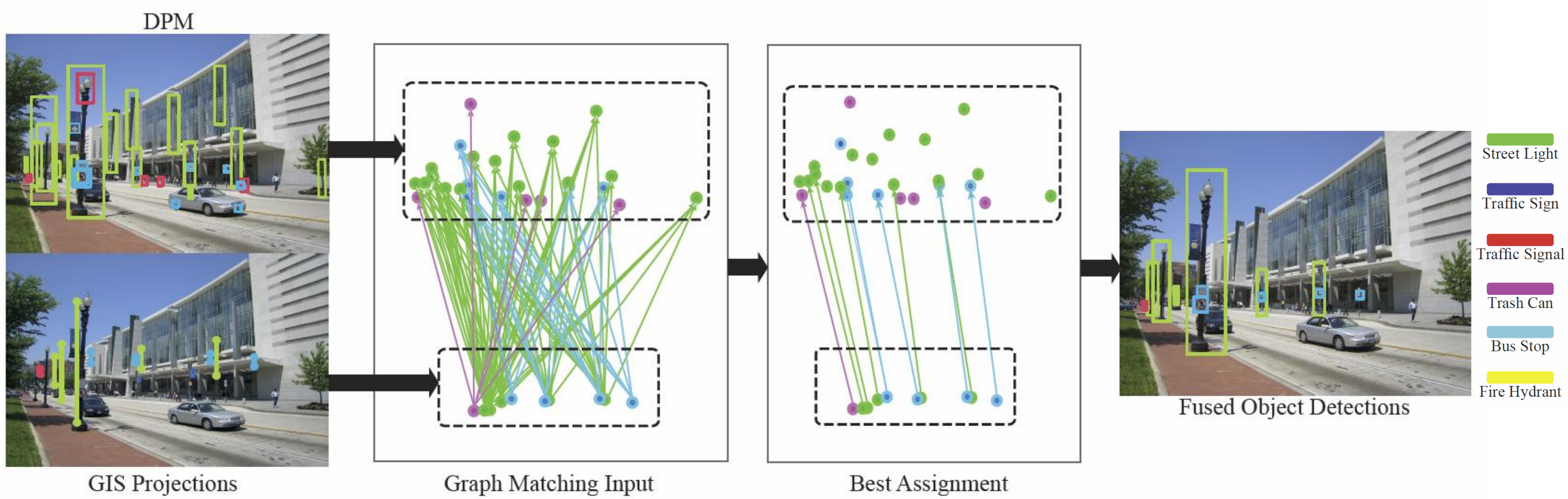
2D projection on the Image

$$\begin{pmatrix} \Phi(i) \\ 1 \end{pmatrix} = \boldsymbol{P}\phi(i)$$

Camera Matrix

GIS priors

Higher Order Graph Matching



New Object Detection Score

$$\mathcal{S}(i) = \alpha \mathcal{S}^G(i) + (1-\alpha)\mathcal{S}^I(i)$$

Affine Fitting Score

DPM Score

$$\mathcal{S}^G(i) = s\left(\boldsymbol{A}\begin{pmatrix}\boldsymbol{\Psi}(i)\\ 1\end{pmatrix} - \begin{pmatrix}\nu(\chi(i))\\ 1\end{pmatrix}\right)$$

Query Image

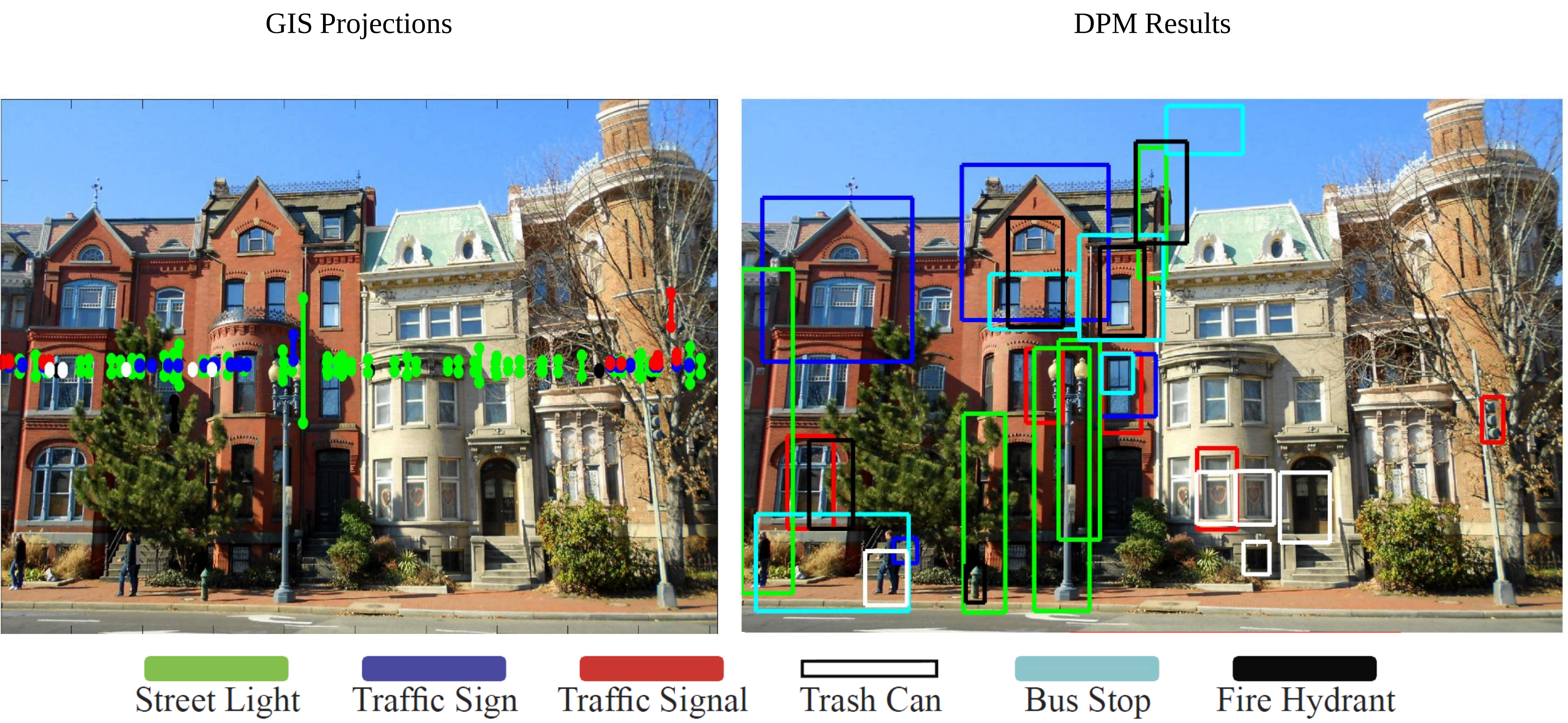
DPM Results

Loose Threshold

Tuned Threshold

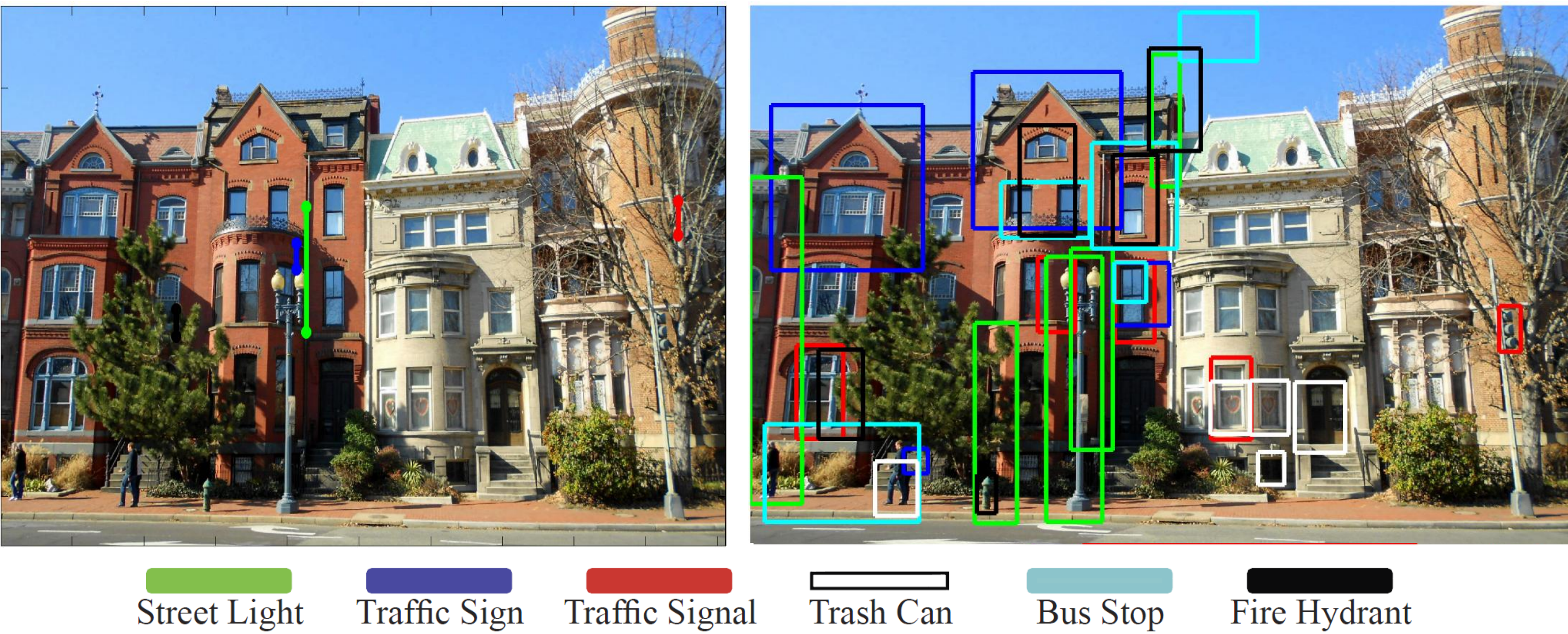
Strict Threshold

Street Light | Traffic Sign | Traffic Signal | Trash Can | Bus Stop | Fire Hydrant


GIS Projections
DPM Results
Street Light
Traffic Sign
Traffic Signal
Trash Can
Bus Stop
Fire Hydrant

Non Occluded GIS Projections

DPM Results

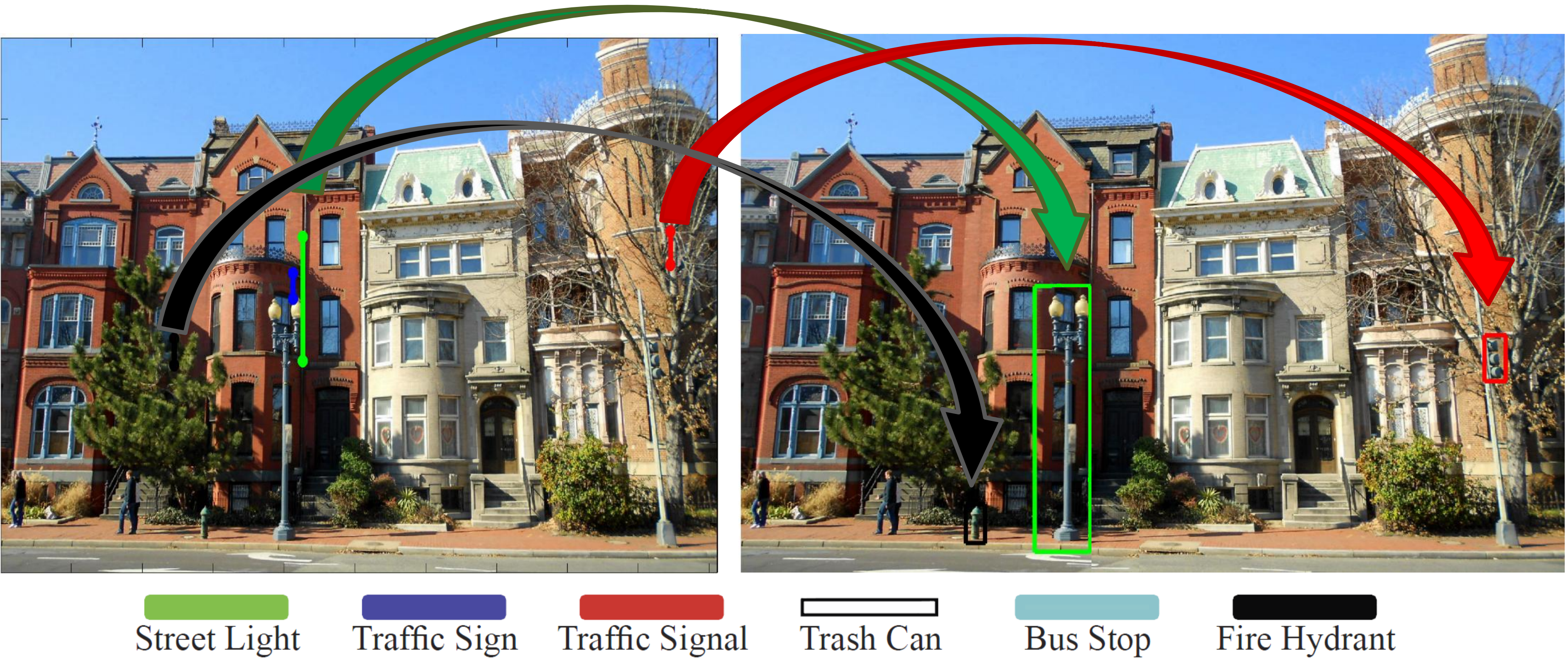



Non Occluded GIS Projections
Our Results
Street Light
Traffic Sign
Traffic Signal
Trash Can
Bus Stop
Fire Hydrant

Traffic Signal, **Street Light**, and **Fire Hydrant** are detected successfully.

Non Occluded GIS Projections

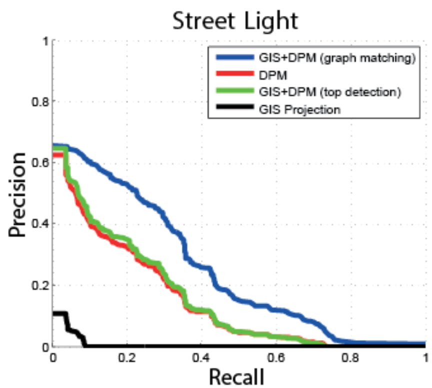
Our Results

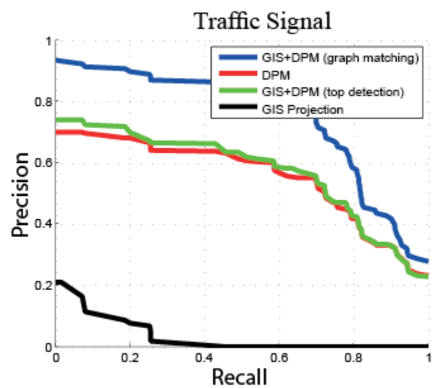


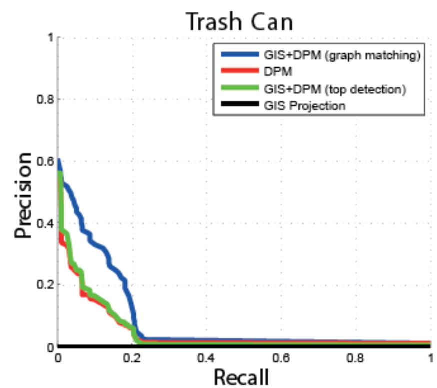
DPM Results (Tuned Threshold)

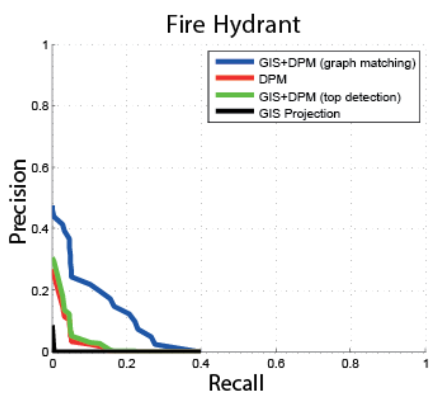
Our Results

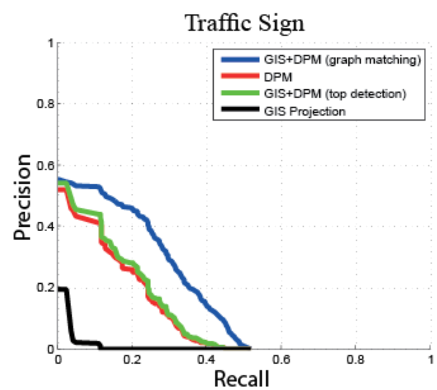
Quantitative Object Detection Results





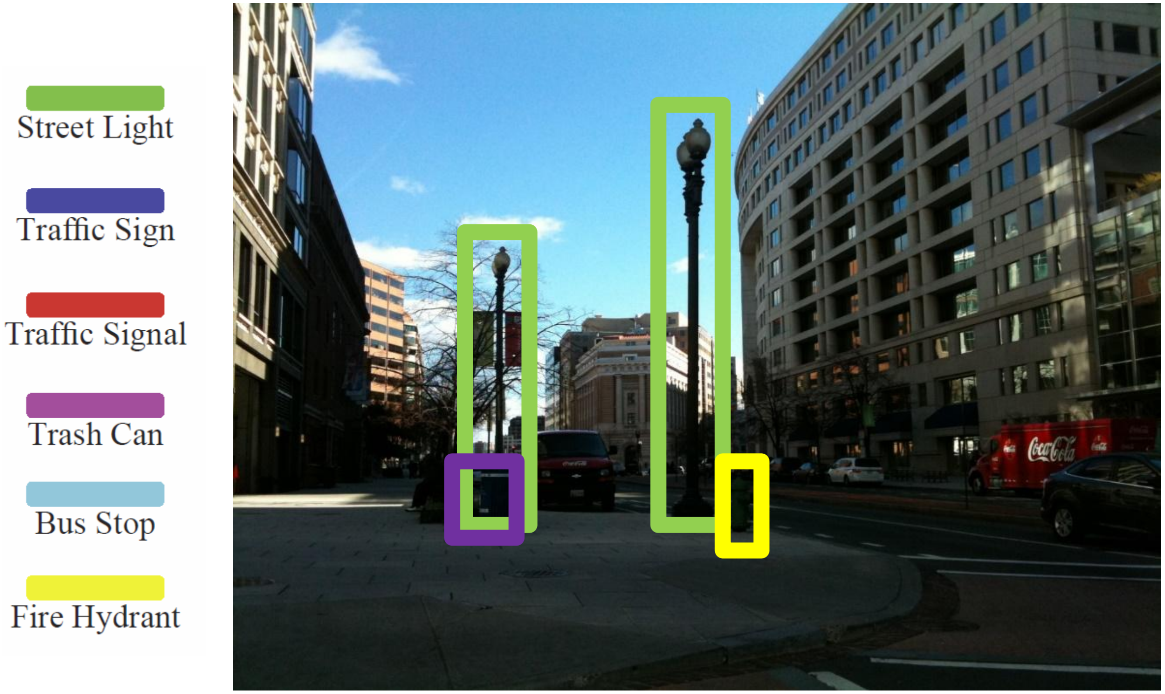




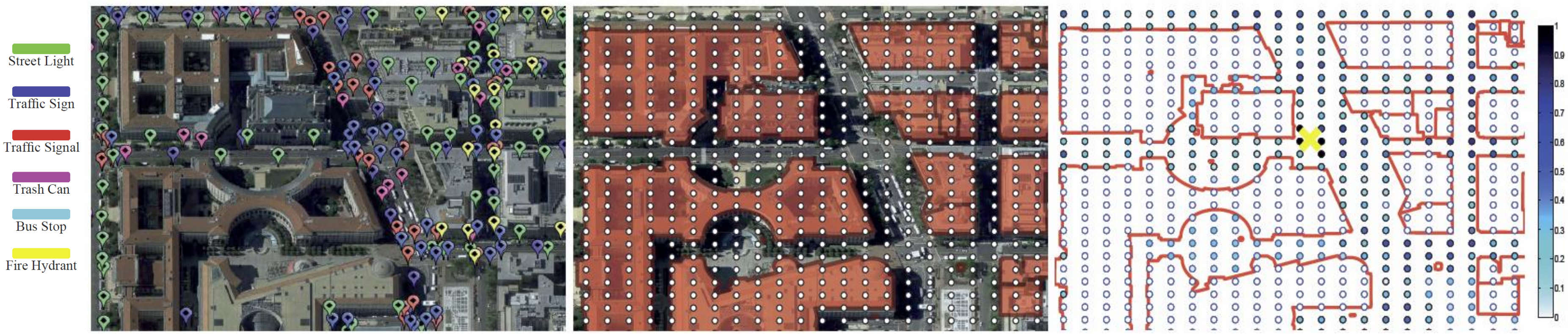


Inverse of this process?!

- Geo-localization using the generic objects?!
  - Cue #1: 2×lamp posts, 1×trash can, 1×fire hydrant.
  - Cue #2: their geometric arrangement.



Geospatial Localization using Generic Objects



Score for Presence of Objects

$$\mathcal{L} = \beta \sum_{i=1}^{|\boldsymbol{\Psi}|} S(i) + (1-\beta) \sum_{c} \frac{\min(|\boldsymbol{\Psi}_c|, |\boldsymbol{\nu}_c|)}{\max(|\boldsymbol{\Psi}_c|, |\boldsymbol{\nu}_c|)}$$

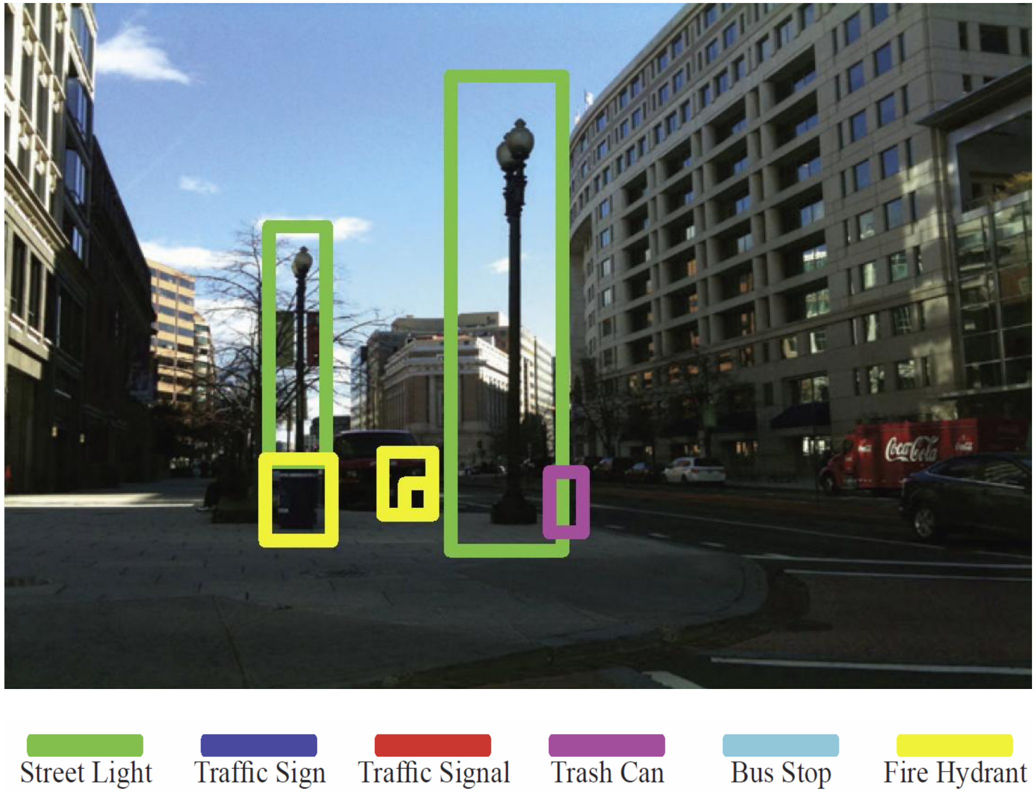
Geometry Preserving Score(Graph Matching)

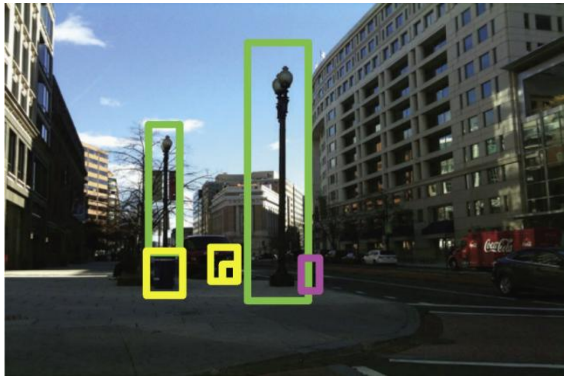
Geospatial Localization

Example 1

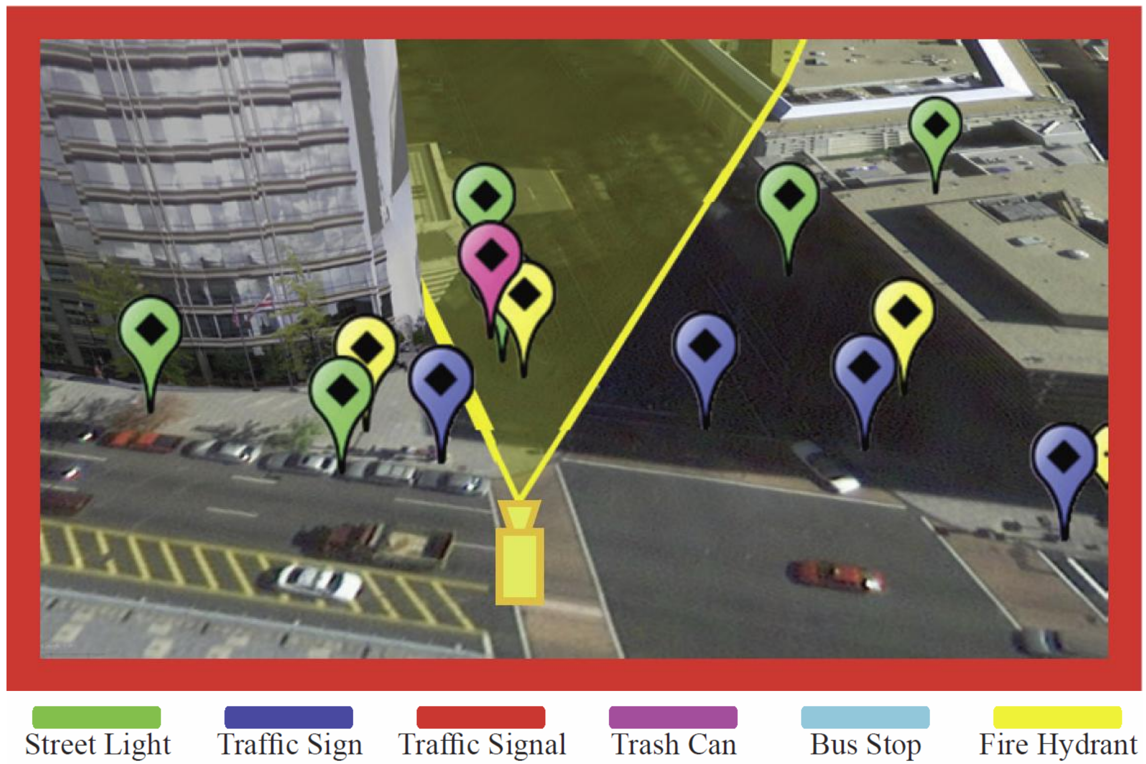
Query Image

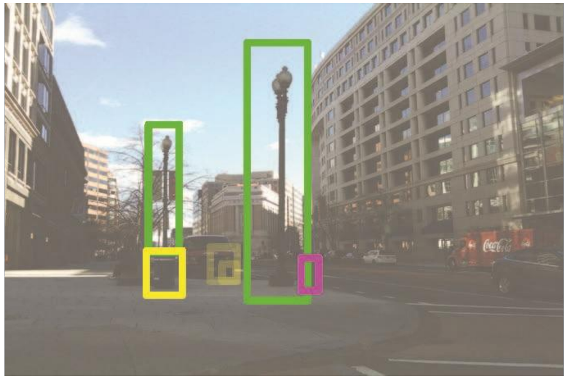
Object Detection Bounding Boxes





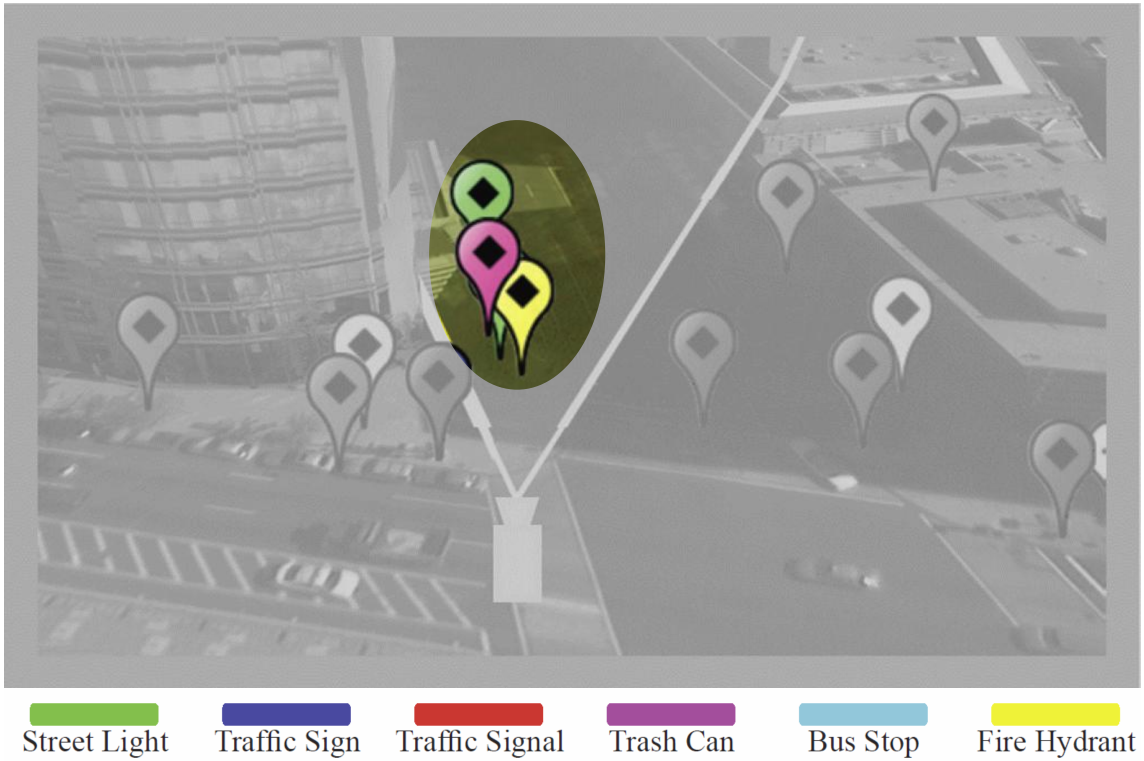
The Correct Location Retrieved as Rank 1

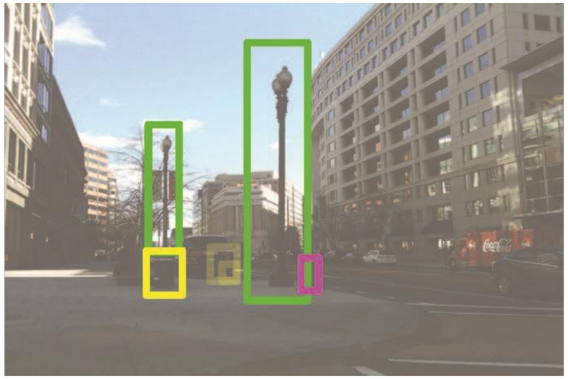




The subset of matching object detections and GIS projections are highlighted.

The Correct Location Retrieved as Rank 1

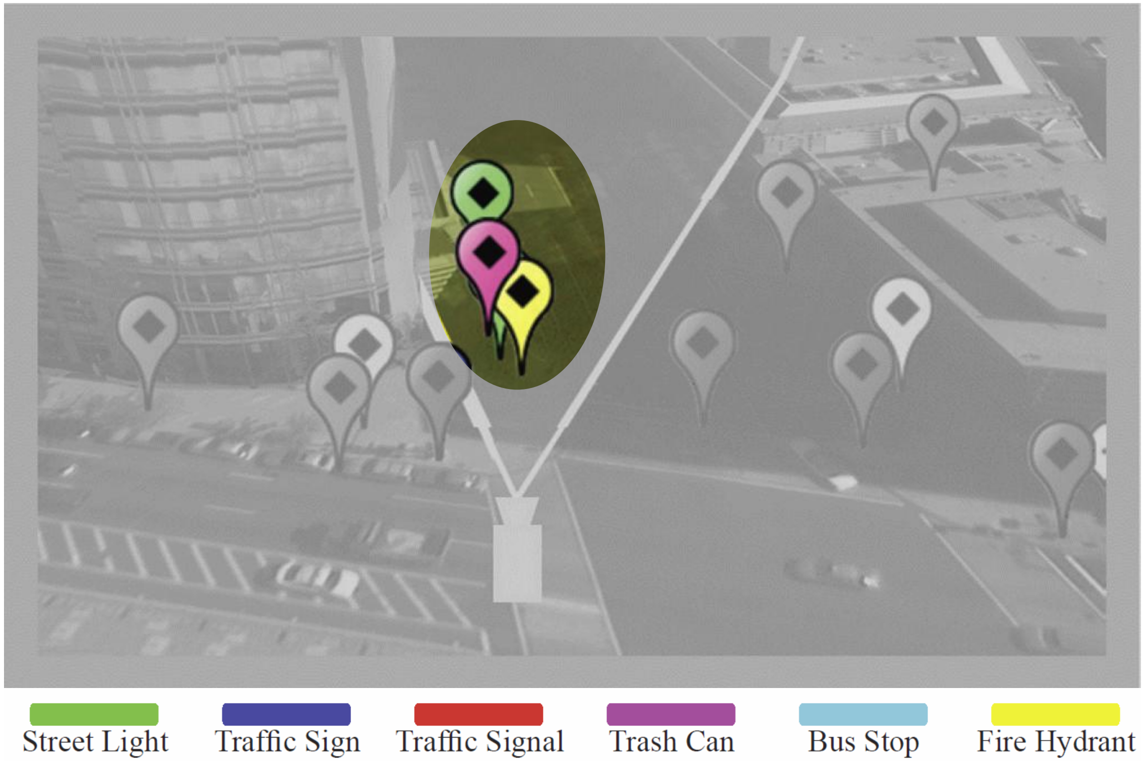




High Score in Presence of Objects: lots of objects in common

High Score in Geometry: Geometric arrangement of objects preserved

The Correct Location Retrieved as Rank 1

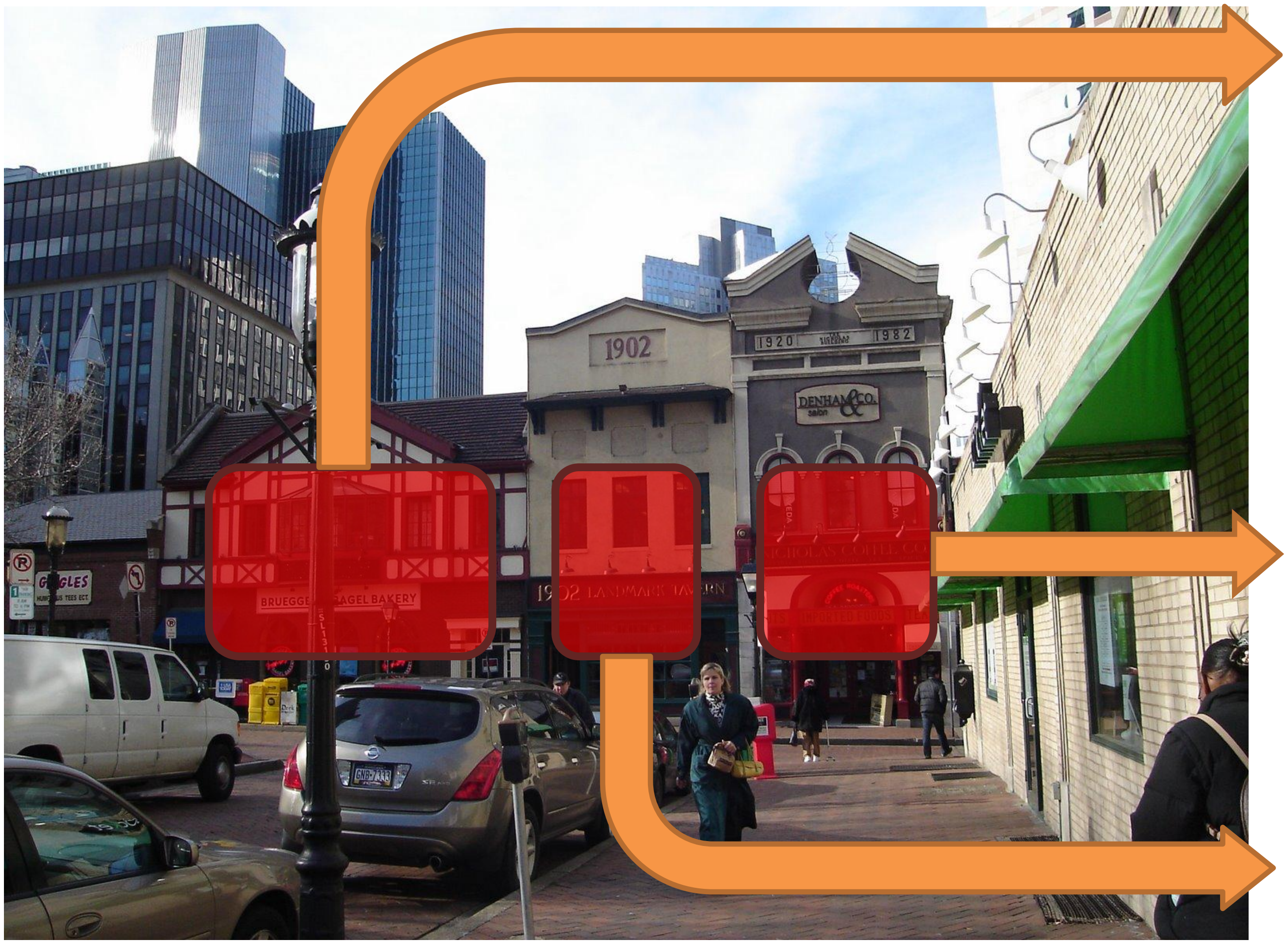


- **Automatic Geo-localization:**
  - Image geo-localization using Generalized Graphs
  - Video Geo-localization and trajectory extraction
- **Robust Refinement of geo-location using Random Walks**
- **Location-Aware image understating:**
  - Location-aware object detection
  - Recognition of storefronts in images

Paper:

Visual Business Recognition - A Multimodal Approach. In ***ACM Multimedia***, 2013.

Visual Business Recognition



Bruegger's Bagel

25 Market Square
Pittsburgh, PA 15222

User Rating: 5/5

Nicholas Coffee Co.

23 Market Square
Pittsburgh, PA 15222

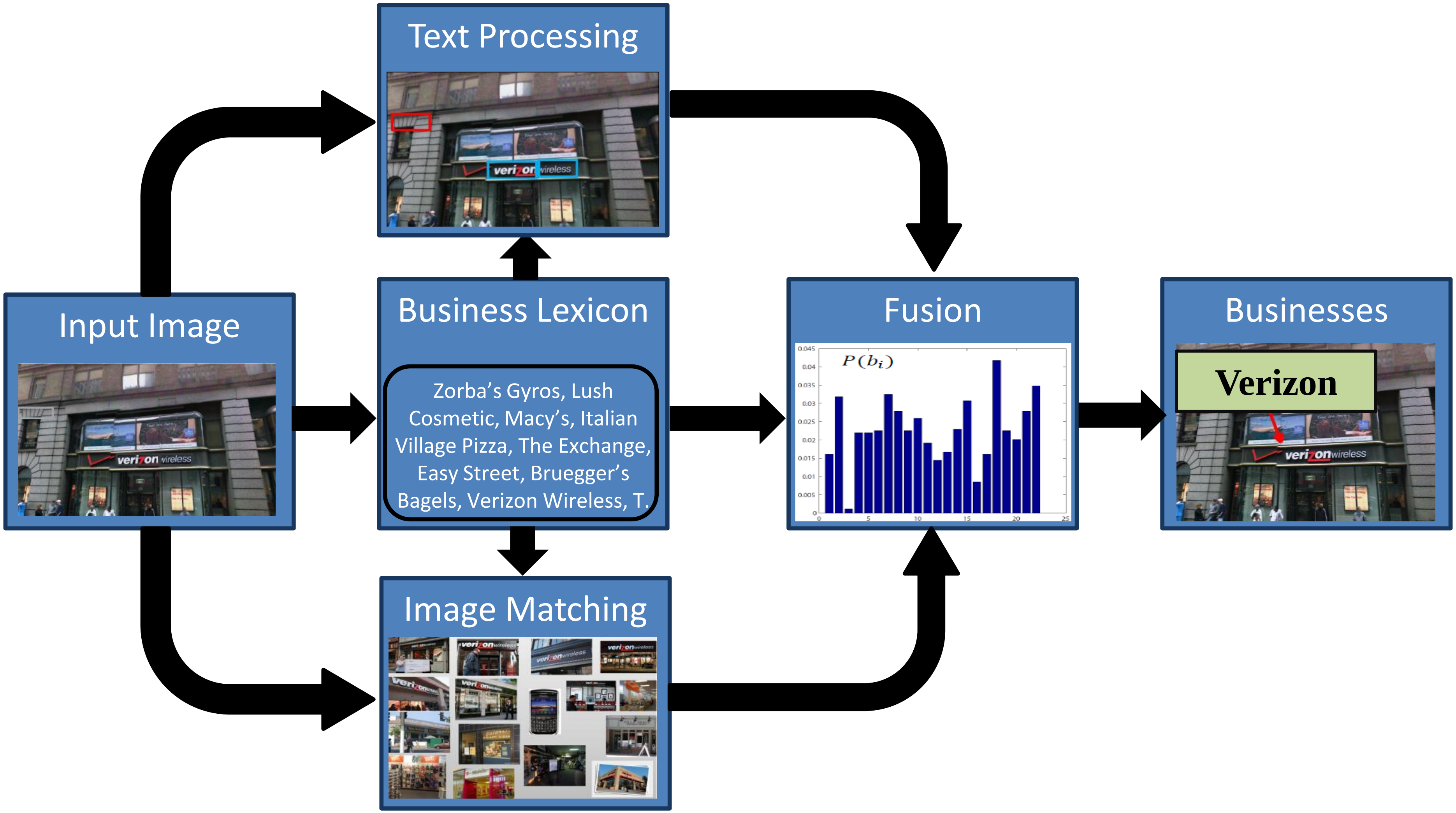
User Rating: 4/5

Tavern,

24 Market Square
Pittsburgh, PA 15222

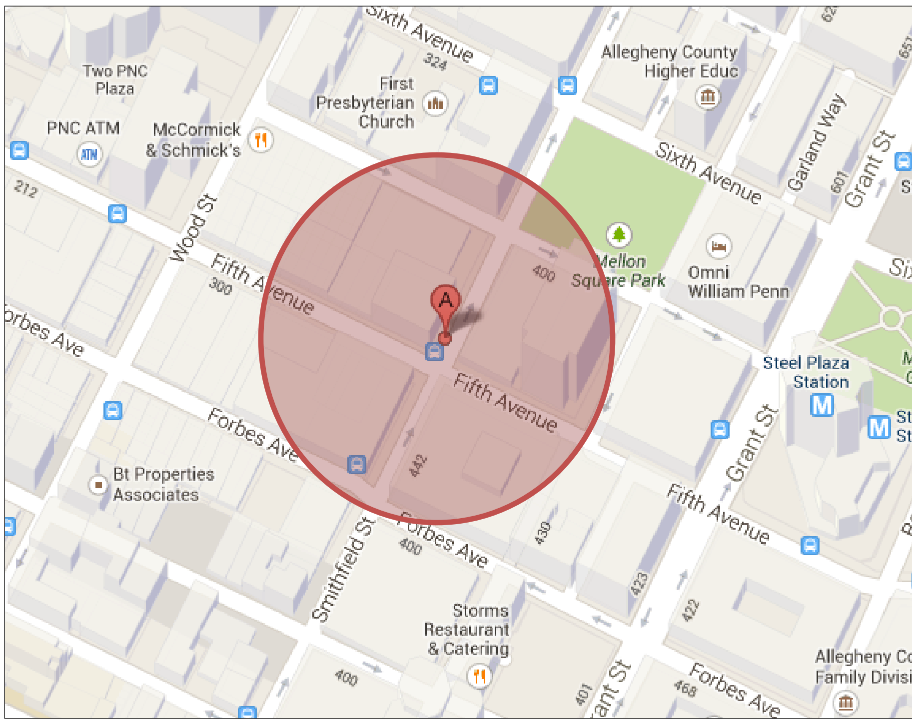
User Rating: 2/5

Block Diagram



Business Lexicon

- An over-complete list of nearby businesses
  - Yelp, Yellow Pages, City Grid, etc.



Gentle Star Medspa, San Francisco Provident Loan Association, Walt Disney Concert Hall, Occasional Occasions, Casa de Campo, Golden Crown Paradise Resort , Great Parnassus Resort & Spa, Ketler Cleaning Service ,Harden Yacht Services, Dicks Auto, The Law Offices of Andrew Gebelt, Will's Handyman, **Verizon Wireless,** Photography Woomer & Hall LLP, Apprehensive Patient Dental Office, Super 8 Motel Sun Prairie 1-877-8-Dump-It Inc, Cvs Pharmacy Ray & Mari's Cleaning Service The Law Offics Of American Wills & Estates, Laurie, Mission St. Advisors, Everlasting, Ayoub Properties…

Text Processing

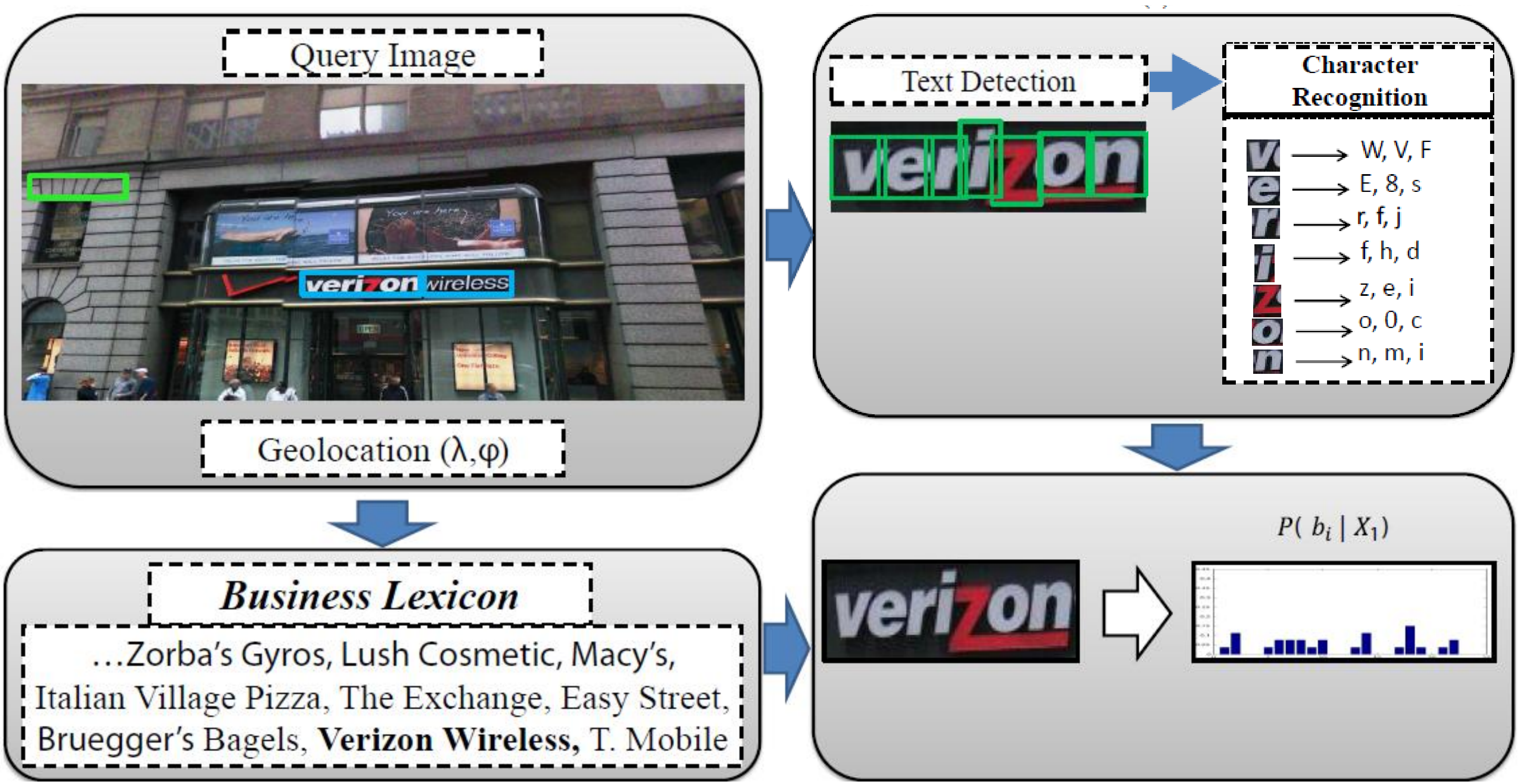


Image Matching

- The image might not have text.
- Too complicated/cluttered text.
- Many ***relevant*** images available on the web

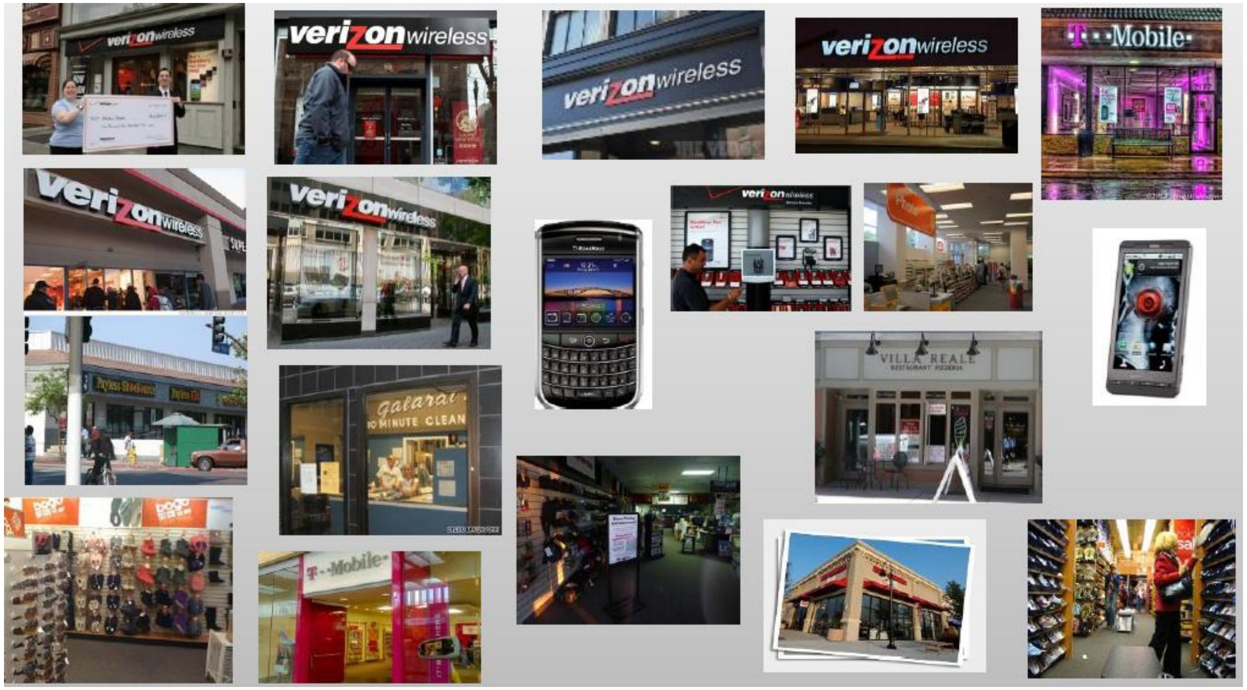


Image Matching

- Business Lexicon as key words
  - "business name", "business name+city", "business name+storefront".

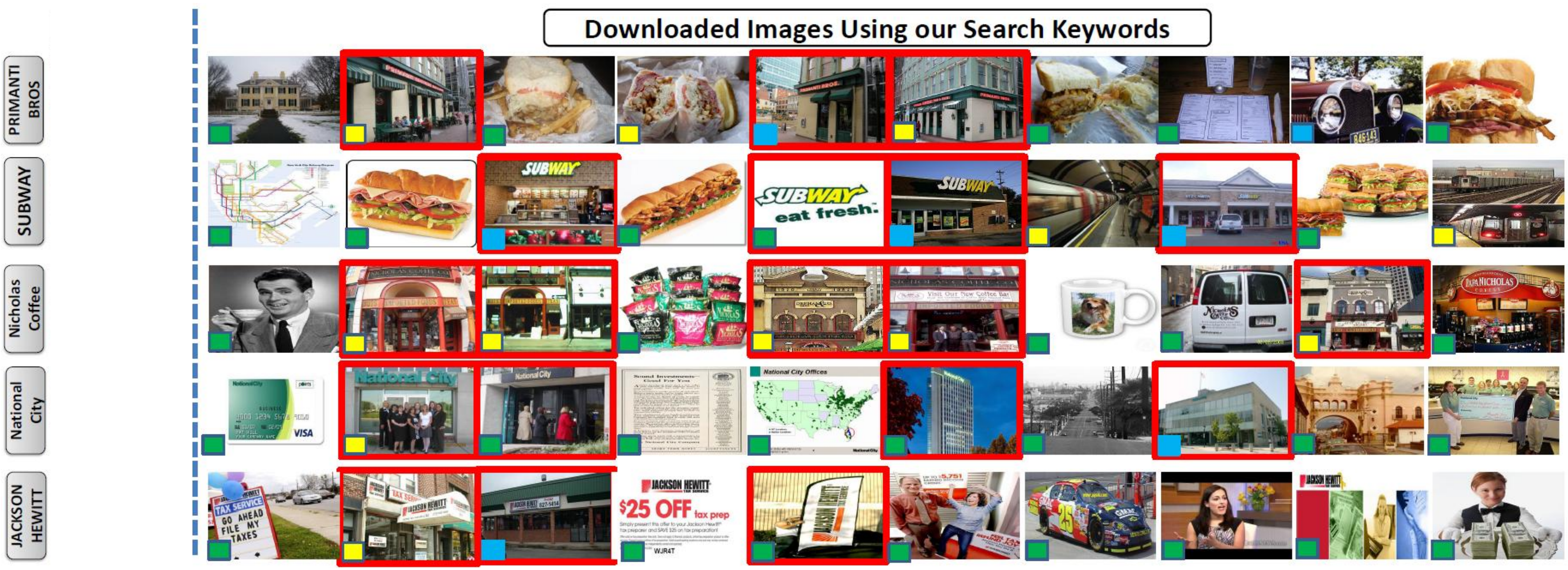
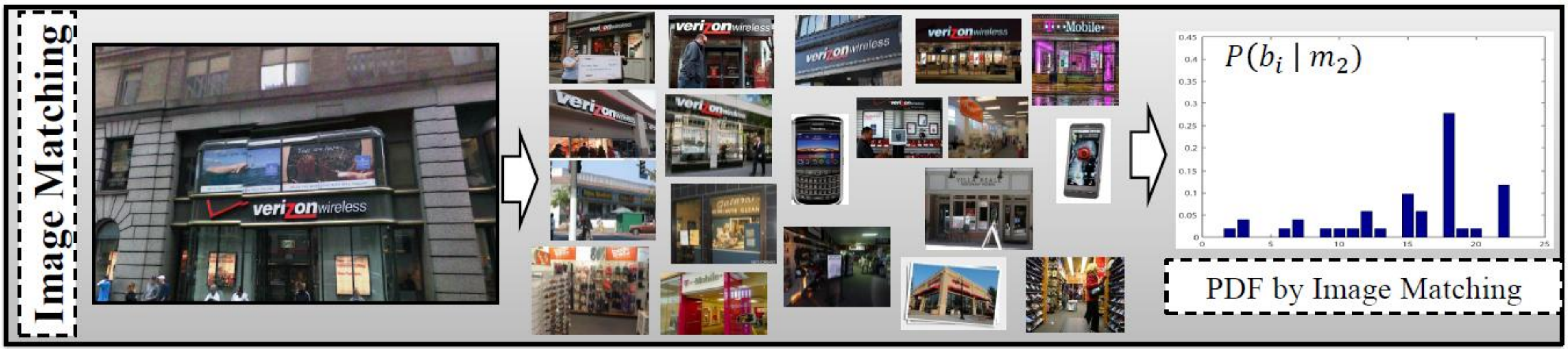


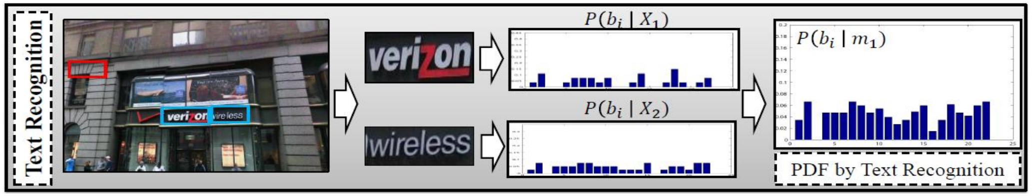
Image Matching

- Between query and web images.
  - Bag of words model.

$$\psi(b_i) = \min_j |h_q - h_{i,j}|, \qquad p(b_i|X) = \frac{sig(\psi(b_i))}{\sum_i sig(\psi(b_i))}$$



Fusing Text Processing and Image Matching

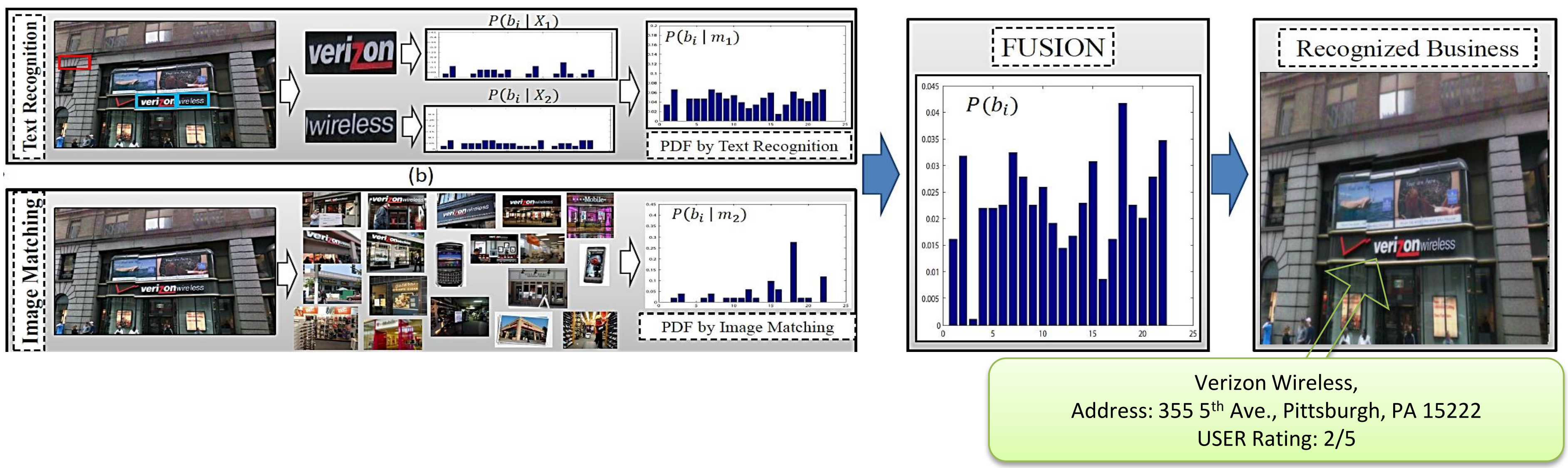


Fusing Text Processing and Image Matching

- Probabilistic Late Fusion

$$p(b_i) = p(b_i|m_1).P(m_1) + p(b_i|m_2).P(m_2)$$

$$P(m_1) = \frac{n_t}{n_t + n_i}, P(m_2) = \frac{n_i}{n_t + n_i}$$

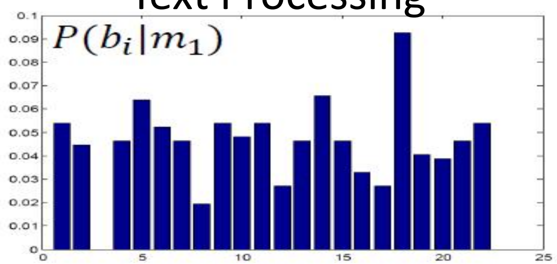


Results

Text Detection

Word Recognition

Text Processing



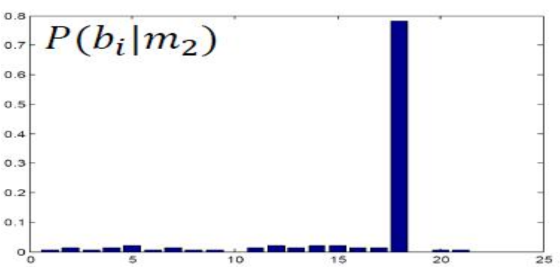
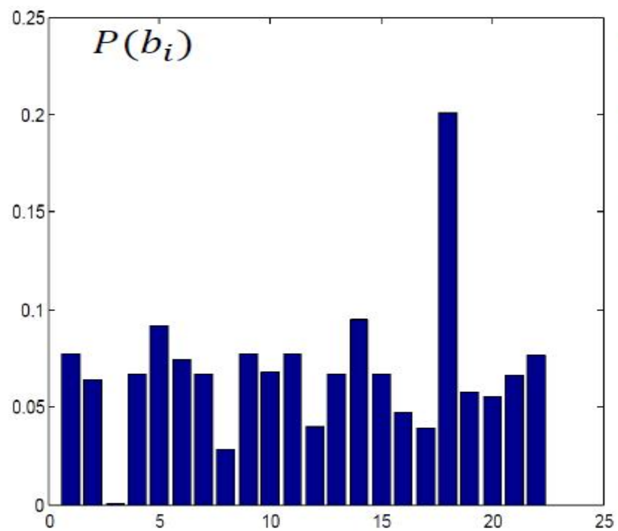


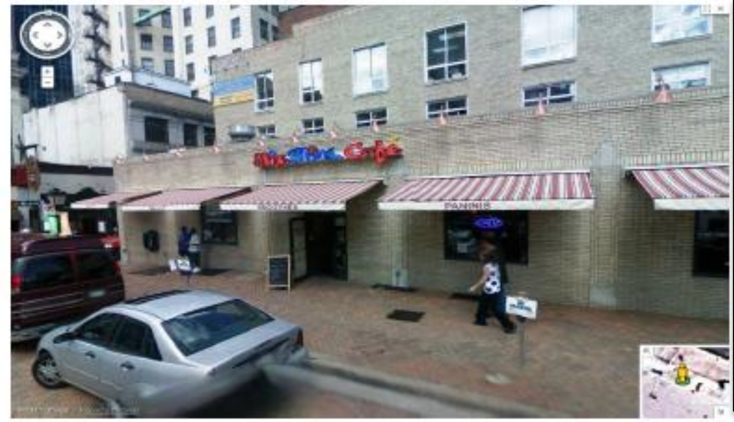
Image Matching

Fusion



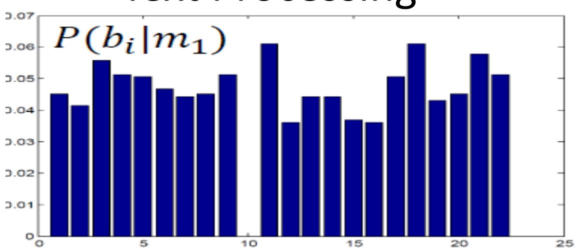
Results

Text Detection



Word Recognition

Text Processing



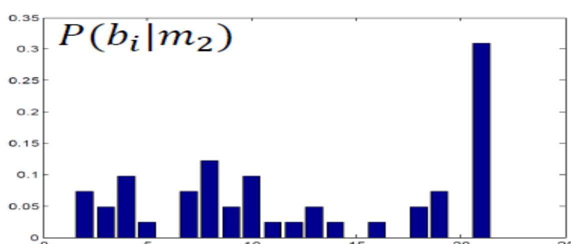
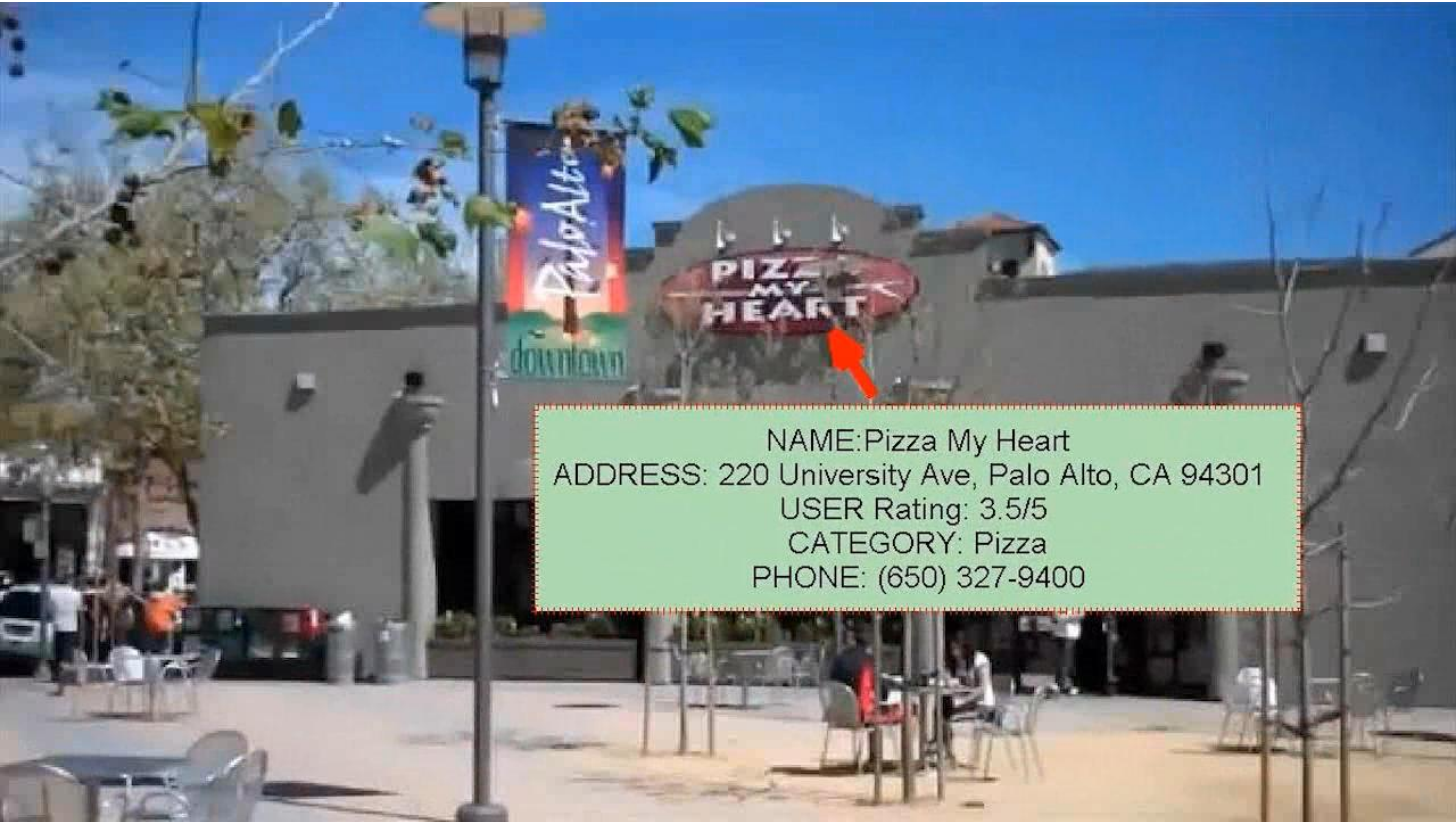


Image Matching

Results


PIZZA MY HEART
Palo Alto
downtown
NAME:Pizza My Heart
ADDRESS: 220 University Ave, Palo Alto, CA 94301
USER Rating: 3.5/5
CATEGORY: Pizza
PHONE: (650) 327-9400

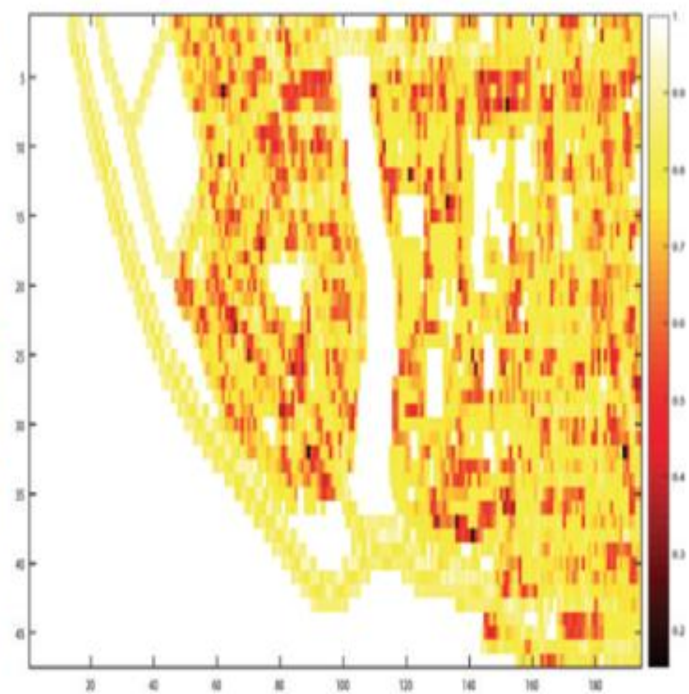
Computational Complexity

- **Image Geo-localization:** sub-linear in search + local feature extraction
- **Video Geo-localization:** sub-linear in search + local feature extraction + Bayesian filtering (matrix multiplication)
- **GPS-Tag Refinement:** sub-linear in search + closed form solution for Random Walks
- **GMCP Matching:** NP-hard, polynomial approximation: $O(KL^2 + LK^\delta)$, L:= # of clusters, K:= # of nodes
- **Object Detection:** DPM + Graph Matching

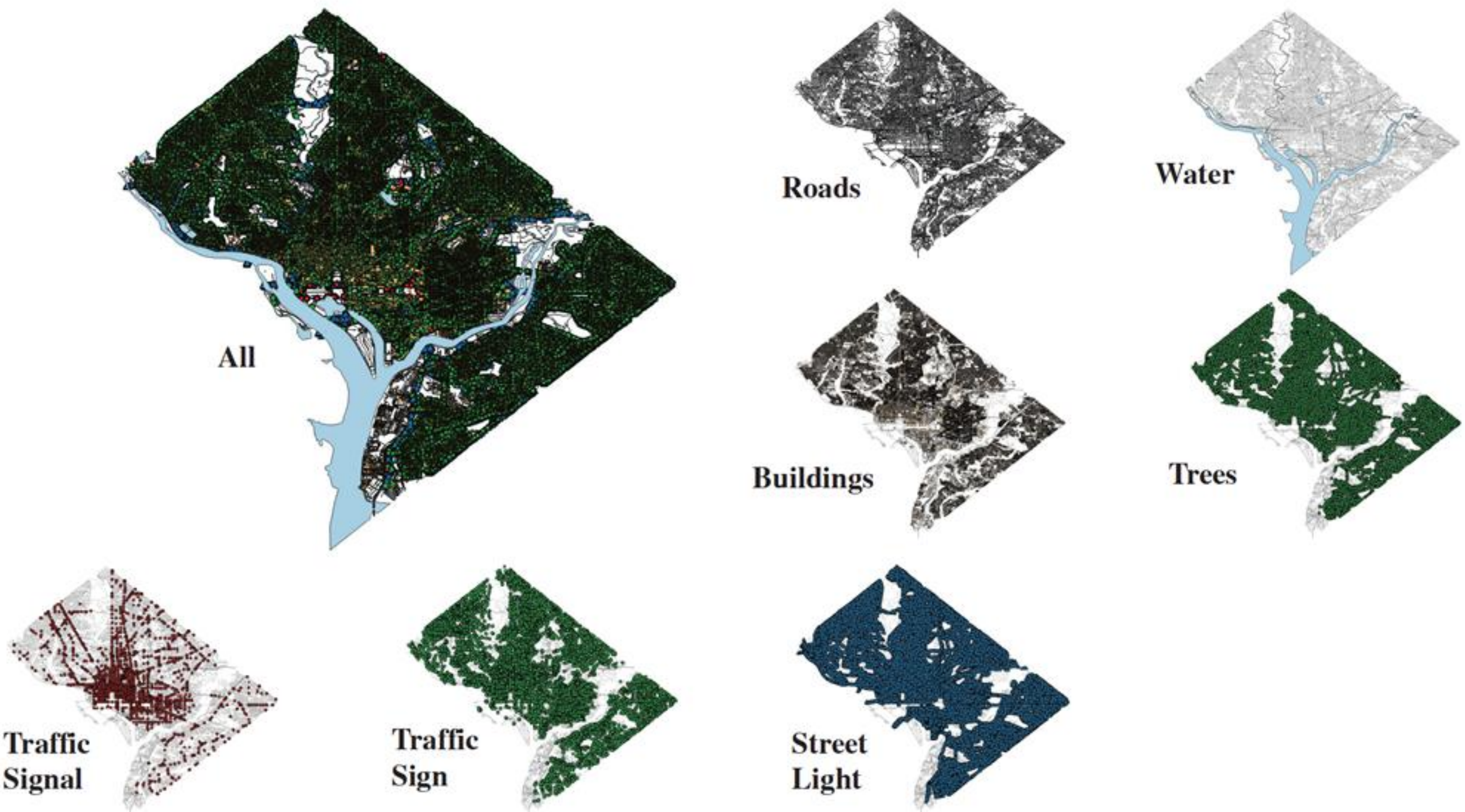
Semantic Cross-View Matching

F. Castalo, A. Zamir, R. Angst, S. Savarese

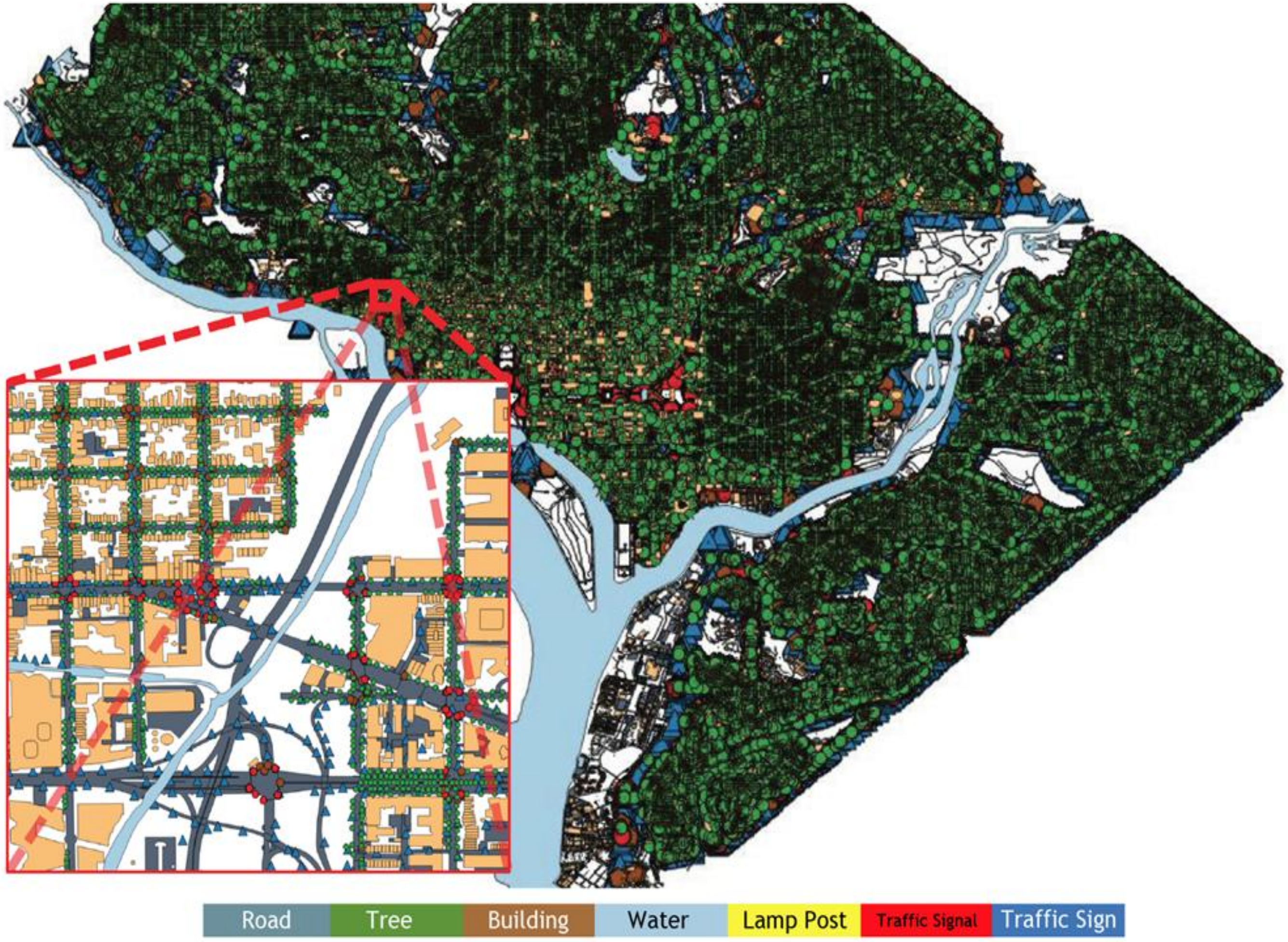
Semantic Wide-Baseline Matching



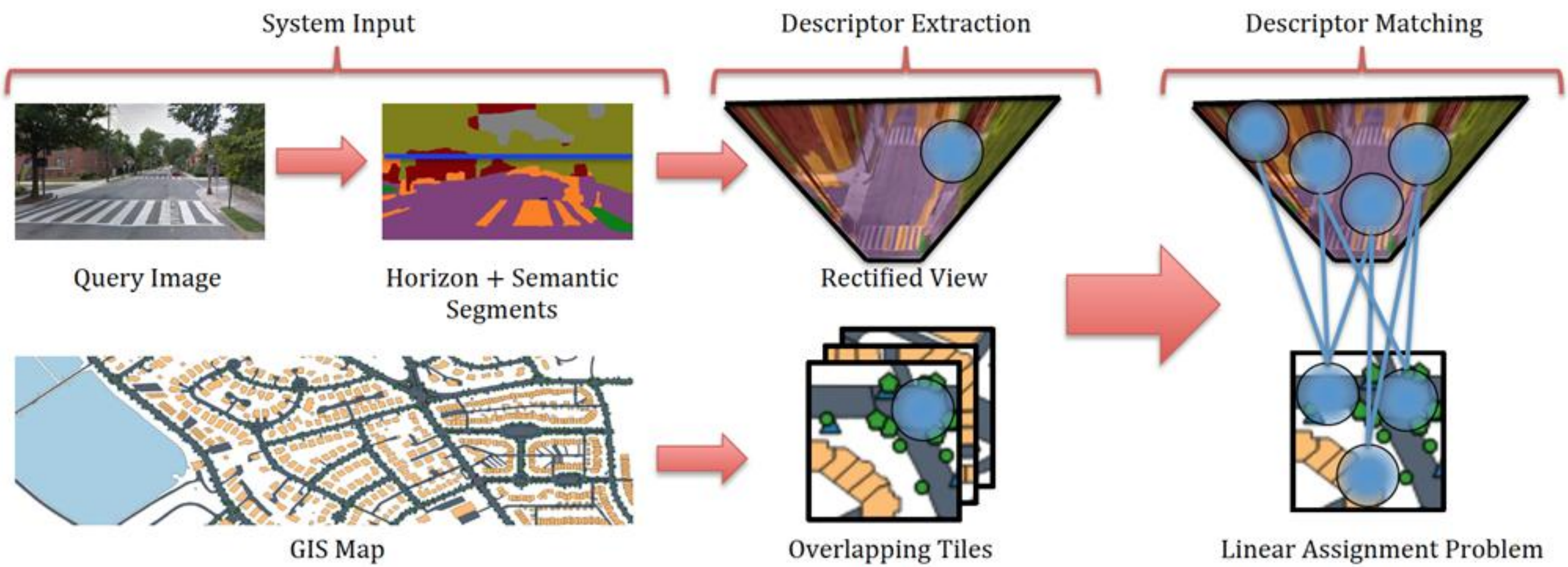
Semantic Map (GIS)



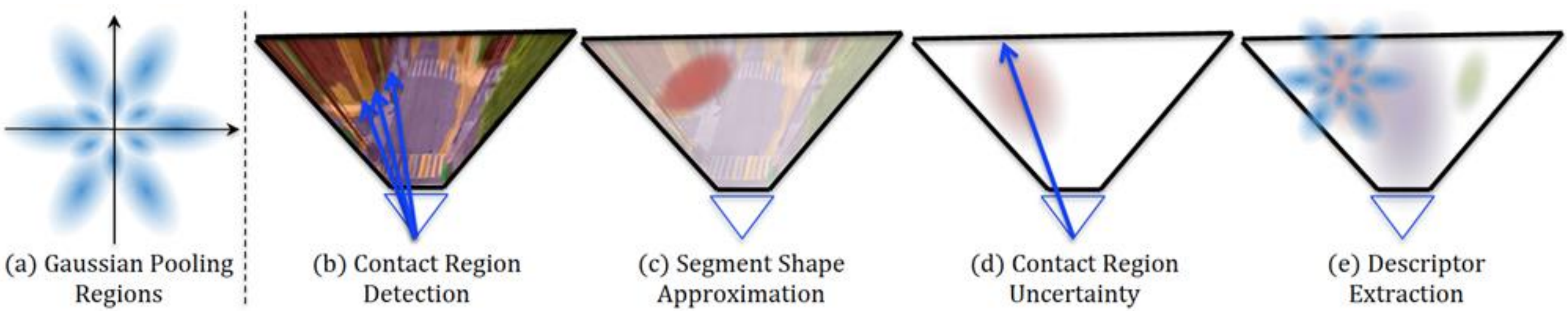
Semantic Map (GIS)

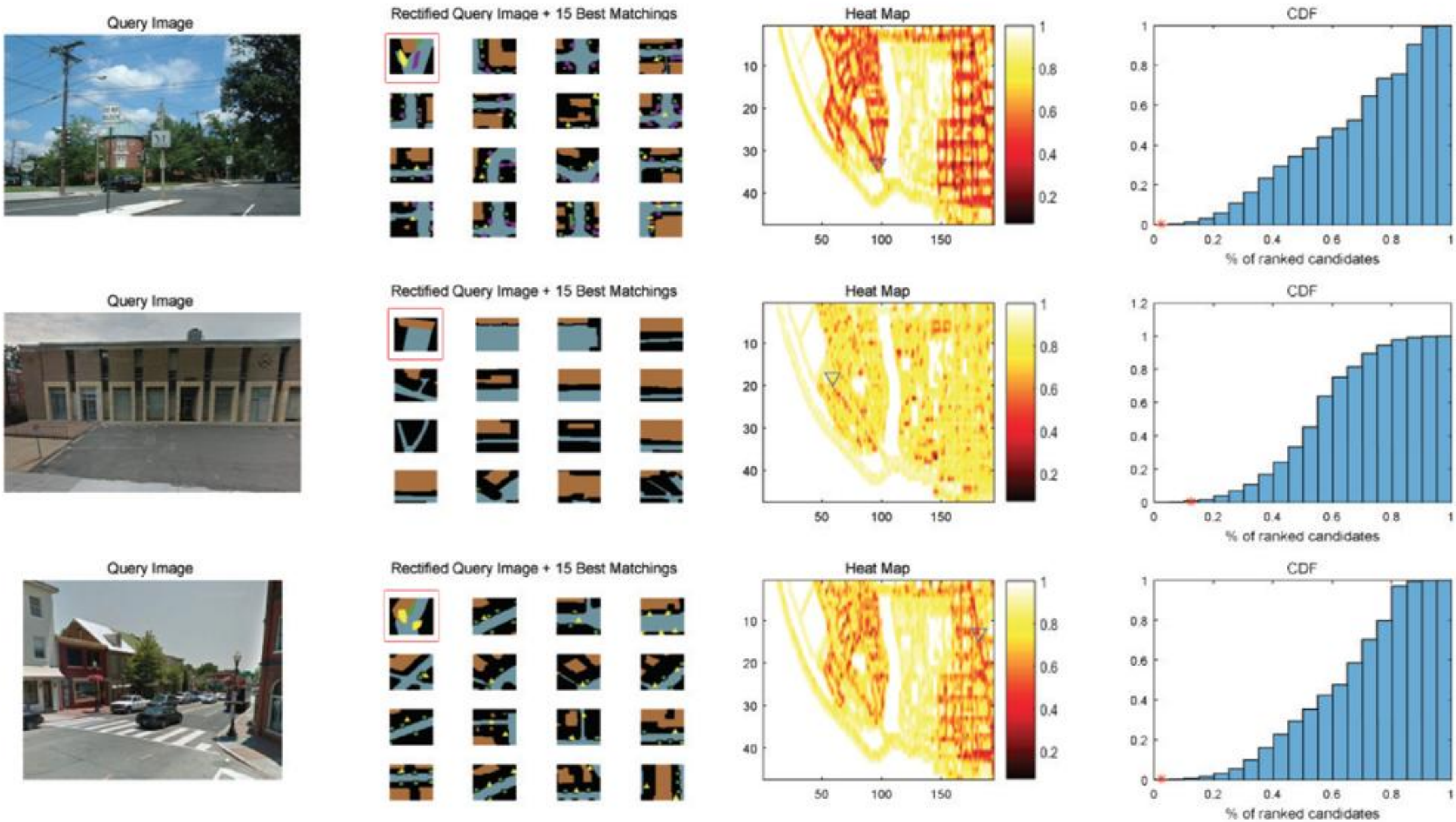


Topological and Semantic Matching

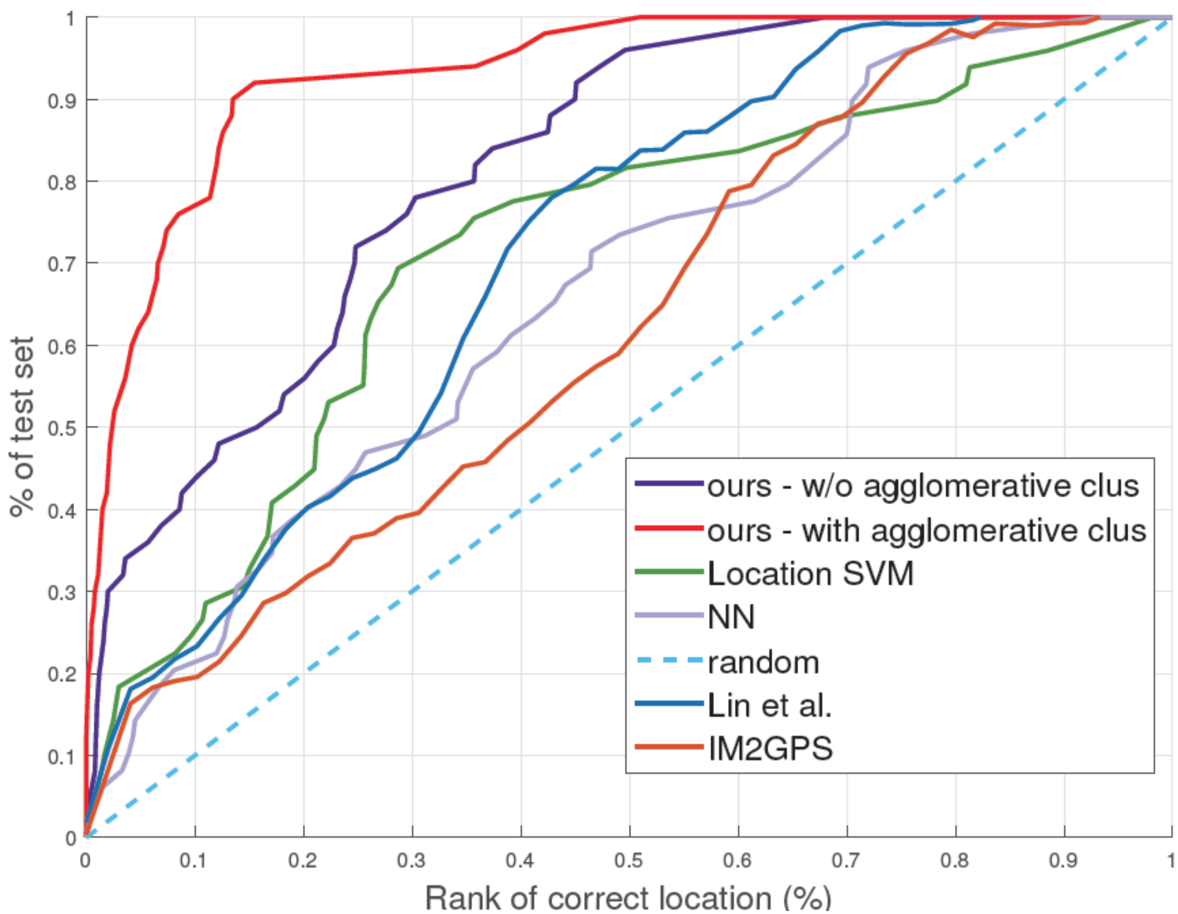


Semantic Segment Layout (SSL) features

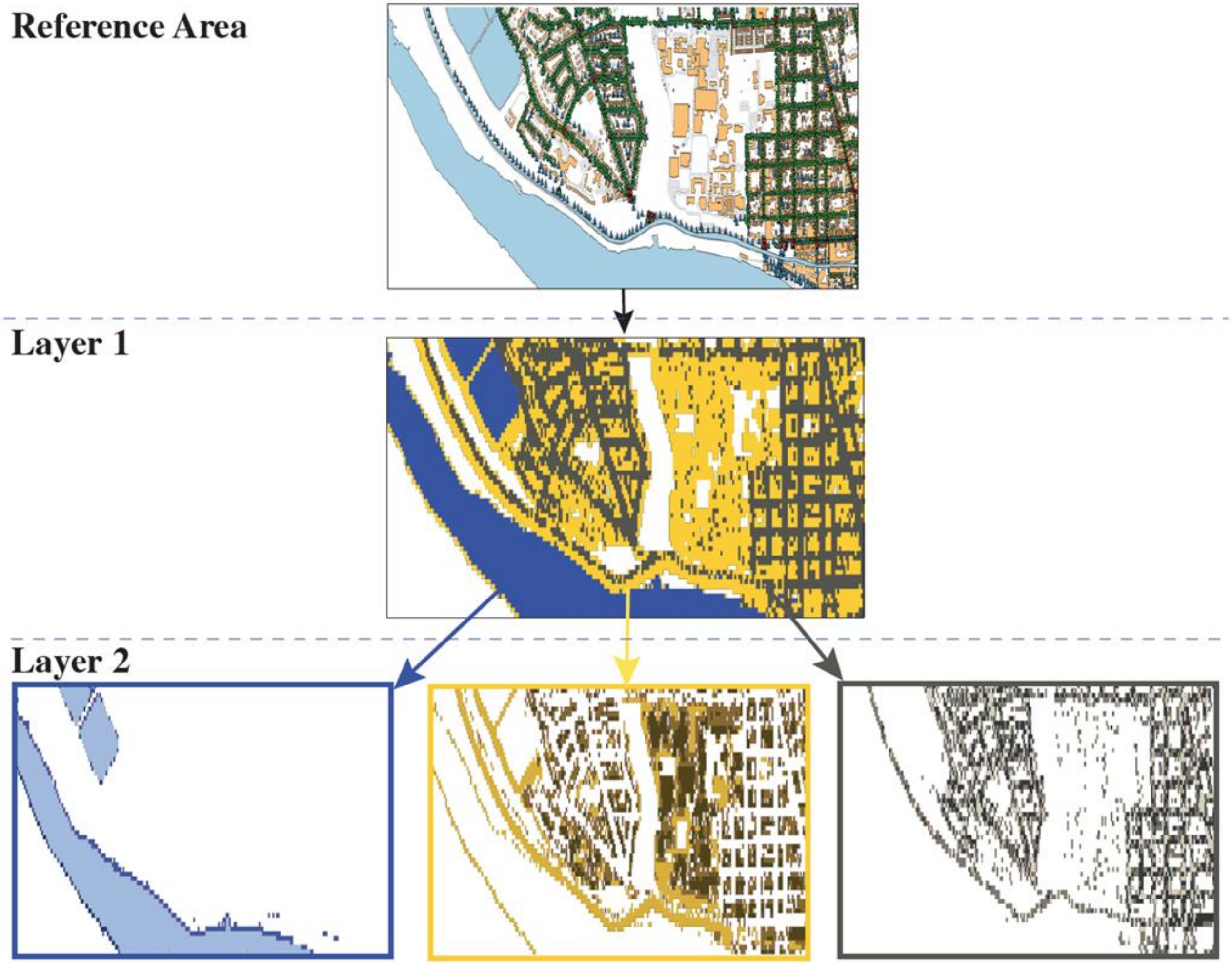



Query Image
Rectified Query Image + 15 Best Matchings
Heat Map
CDF
% of ranked candidates
Query Image
Rectified Query Image + 15 Best Matchings
Heat Map
CDF
% of ranked candidates
Query Image
Rectified Query Image + 15 Best Matchings
Heat Map
CDF
% of ranked candidates

Experimental Results



Semantic Tree



Thank You!

zamir@cs.stanford.edu

A. Zamir, F. Castaldo, R. Angst, S. Savarese, G. Vaca, S. Ardeshir, M. Shah

- **PAMI'14:** Image Geo-localization Based on Multiple-NN Feature Matching Using Generalized Graphs
- **CVPR'14:** GPS-Tag Refinement using Random Walks with an Adaptive Damping Factor
- **ECCV'14:** GIS-Assisted Object Detection and Geospatial Localization
- **ACM Multimedia'13:** Visual Business Recognition - A Multimodal Approach
- **CVPR'12:** City Scale Geo-spatial Trajectory Estimation of a Moving Camera
- **ICMLA'11:** Identification of Commercial Entities in Street View Imagery
- **ECCV'10:** Accurate Image Localization Based on Google Maps Street View
- **In Submission'15:** Semantic Cross-View Matching