Large Scale Cross View Image Geo-localization

Dr. Chen Chen
Department of Electrical and Computer Engineering
E-mail: chen.chen@uncc.edu
https://webpages.uncc.edu/cchen62/
What is image geo-localization?

Input

Visual Information (Images)

Output

Location in terms of Longitude and Latitude

40.4419, -79.9986
What is image geo-localization?

Query street-view image

Geo-tagged reference database

Find match

Street-view images (i.e., same view)

GPS location?

(Latitude, Longitude) = (40.441426, -80.003586)

Top-1 match (ranked by similarity)
What is image geo-localization?

Query street-view image

Find match

Aerial images (i.e., cross view)

Geo-tagged reference database

GPS location?

(Latitude, Longitude) = (40.441426, -80.003586)
Why is image geo-localization important?

- Accurate Visual Localization for Automotive Applications

Why is image geo-localization important?

- Cross-View Policy Learning for Street Navigation

Why is image geo-localization important?

- UAV Pose Estimation using Cross-view Geo-localization

Street-to-Aerial View Matching for Image Geo-localization and Orientation Estimation


Main Challenges

- Appearance Gap

- Sample Imbalance
  - The number of positive samples for an anchor street-view image is very limited in geo-localization, i.e., only one.
Main Challenges

- Image alignment

<table>
<thead>
<tr>
<th>Validation</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aligned</td>
</tr>
<tr>
<td>Aligned</td>
<td>60.1%</td>
</tr>
<tr>
<td>Rotate</td>
<td>13.5%</td>
</tr>
</tbody>
</table>

Top-1 recall accuracy with different alignment settings
Overall Framework (network architecture)
Matching Loss

Binomial deviance loss (Yi et al.)

\[
L = \frac{1}{N_p} \sum_{i}^{N_p} \sigma(-\alpha(s_i^p - m)) + \frac{1}{N_n} \sum_{i}^{N_n} \sigma(\alpha(s_i^n - m))
\]

\(s_i^p\) and \(s_i^n\) denote the cosine similarity between the i-th anchor and its positive and negative samples

\(N_p\) and \(N_n\) represent the number of positive and negative pairs

\(m\) : a positive margin parameter

Matching Loss

Our new loss function

\[ L = \sum_{i}^{N_p} \frac{\sigma(-\alpha_p (s_i^p - m_p))}{\alpha_p N_p} + \sum_{i}^{N_n} \frac{\sigma(\alpha_n (s_i^n - m_n))}{\alpha_n N_n} \]

When positive samples are much fewer than negative samples, as in cross-view geo-localization with only one positive match, it would be easier to pulling the only matched sample close to the anchor rather than pushing all negative samples away (i.e., assign a much smaller value to \( \alpha_p \) than \( \alpha_n \)).
## Geo-localization Results

<table>
<thead>
<tr>
<th>Method</th>
<th>CVUSA</th>
<th></th>
<th>Vo</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1%</td>
<td>Top-1</td>
<td>Top-1%</td>
<td>Top-1</td>
</tr>
<tr>
<td>Scott [22] (ICCV’15)</td>
<td>34.3%</td>
<td>–</td>
<td>15.4%</td>
<td>–</td>
</tr>
<tr>
<td>Zhai [26] (CVPR’17)</td>
<td>43.2%</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Vo [21] (ECCV’16)</td>
<td>63.7%</td>
<td>–</td>
<td>59.9%</td>
<td>–</td>
</tr>
<tr>
<td>CVMNet [8] (CVPR’18)</td>
<td>93.6%</td>
<td>22.5%</td>
<td>67.9%</td>
<td>–</td>
</tr>
<tr>
<td>Lending [13] (CVPR’19)</td>
<td>93.19%</td>
<td>31.71%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Reweight [3] (ICCV’19)</td>
<td>98.3%</td>
<td>46.0%</td>
<td>78.3%</td>
<td>–</td>
</tr>
<tr>
<td>GAN [14] (ICCV’19)</td>
<td>95.98%</td>
<td>48.75%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>97.7%</strong></td>
<td><strong>54.5%</strong></td>
<td><strong>88.3%</strong></td>
<td><strong>11.8%</strong></td>
</tr>
</tbody>
</table>

Table 2: Top-1 and top-1% recall accuracy comparison on CVUSA and Vo datasets.
Geo-localization Examples

Street view id: 6312, file name: 0023612.jpg
Geo-localization Examples

Street view id: 5134, file name: 0037767.jpg

Aerial view rank: 5

Top-1

Top-2

Top-3

Top-4

Top-5
Geo-localization Examples

Street view id: 18110

Aerial view rank: 1
Similarity: 0.76

Top-1
Similarity: 0.76

Top-2
Similarity: 0.75

Top-3
Similarity: 0.74
Geo-localization Examples

A failure case
Visual Explanation of the Matching Results

• Visual explanation using Grad-CAM
What is Grad-CAM?

- Gradient-weighted Class Activation Mapping (Grad-CAM)

Visual Explanation of the Matching Results

The most activated regions are likely to be the same objects.
Visual Explanation of the Matching Results

Street view id: 49288

Aerial view positive, id: 49288

Similarity: 0.74
An Interesting Finding

We find the Grad-CAM activation maps have the rotation-invariant property!
Orientation Estimation with Grad-CAM

We find the Grad-CAM activation maps have the rotation-invariant property!
The angle distributions of activated pixels from two views would be similar if the image pair is well aligned. Find the angle $\phi$ so that $p_{aerial}(\theta + \phi)$ best matches $p_{street}(\theta)$. The confidence is calculated as $\frac{0.06}{0.06 + 0.02} = 0.75$. The peak values are as follows: Peak1: 0.06 and Peak2: 0.02.
Orientation Estimation Example

$126^\circ$

$125.2^\circ$

Conf = 0.95
Summary

• Ablation study and visual explanation lead a key observation – the alignment has a great impact on the retrieval performance

• We show that improvements on metric learning techniques boost the retrieval performance

• We discover that the orientation information between cross-view images can be estimated when the alignment is unknown
VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval

One-to-one Retrieval

• Existing works simply assume that each query ground-view image has one corresponding reference aerial-view image whose center is exactly aligned at the location of the query image.

• This is not practical for real-world applications, because the query image may be generated at arbitrary locations in the area of interest and the reference images should be captured before the queries emerge.
VIGOR

Dataset Setting: given an area of interest (AOI), the reference aerial images are densely sampled to achieve a seamless coverage of the AOI and the street-view queries are captured at arbitrary locations.
Data Distribution

Manhattan  Chicago  San Francisco  Seattle
## Datasets Comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Satellite images</td>
<td>~450,000</td>
<td>128,334</td>
<td>44,416</td>
<td>90,618</td>
</tr>
<tr>
<td>Panoramas in total</td>
<td>~450,000</td>
<td>128,334</td>
<td>44,416</td>
<td>238,696</td>
</tr>
<tr>
<td>Panoramas after balancing</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>105,214</td>
</tr>
<tr>
<td>Street-view GPS locations</td>
<td>Aligned</td>
<td>Aligned</td>
<td>Aligned</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Full panorama</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Multiple cities</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Orientation information</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Evaluation in terms of meters</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Seamless coverage on area of interest</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Number of references covering each query</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
Example Query and Reference
Coarse-to-fine Cross-view Localization

Beyond One-to-one

How to make use of the semi-positive images?

Directly considering semi-positive as positive results in a low accuracy. We force the ratio of the similarities in the embedding space to be close to the ratio of IOUs.

IOU-based semi-positive assignment loss

\[ L_{IOU} = \left( \frac{S_{semi}}{S_{pos}} - \frac{IOU_{semi}}{IOU_{pos}} \right)^2 \]

<table>
<thead>
<tr>
<th>Semi-positive Assignment</th>
<th>Same-Area</th>
<th>Cross-Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>No semi-positive (i.e. baseline, ( L_{triplet} ))</td>
<td>38.0</td>
<td>62.9</td>
</tr>
<tr>
<td>Positive (( L_{triplet} ))</td>
<td>20.3</td>
<td>45.7</td>
</tr>
<tr>
<td>IOU (( L_{triplet} + L_{IOU} ))</td>
<td><strong>41.1</strong></td>
<td><strong>65.9</strong></td>
</tr>
</tbody>
</table>
Beyond Retrieval

\[ \mathcal{L}_{\text{offset}} = (\text{lat} - \text{lat}^*)^2 + (\text{lon} - \text{lon}^*)^2 \]

\text{lat} and \text{lon} denote the predicted offset of the query GPS location relative to the reference GPS. \text{lat}^* and \text{lon}^* denote the ground-truth offset.
Comparison with State-of-the-art

- Retrieval Performance

<table>
<thead>
<tr>
<th></th>
<th>Same-Area</th>
<th></th>
<th></th>
<th>Cross-Area</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1%</td>
<td>Hit Rate</td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>Siamese-VGG ($L_{triplet}$)</td>
<td>18.1</td>
<td>42.5</td>
<td>97.5</td>
<td>21.2</td>
<td>2.7</td>
<td>8.2</td>
</tr>
<tr>
<td>SAFA ($L_{triplet}$)</td>
<td>33.9</td>
<td>58.4</td>
<td>98.2</td>
<td>36.9</td>
<td>8.2</td>
<td>19.6</td>
</tr>
<tr>
<td>SAFA+Mining (baseline, $L_{triplet}$)</td>
<td>38.0</td>
<td>62.9</td>
<td>97.6</td>
<td>41.8</td>
<td>9.2</td>
<td>21.1</td>
</tr>
<tr>
<td>Ours ($L_{hybrid}$)</td>
<td><strong>41.1</strong></td>
<td><strong>65.8</strong></td>
<td><strong>98.4</strong></td>
<td><strong>44.7</strong></td>
<td><strong>11.0</strong></td>
<td><strong>23.6</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Top-1</th>
<th>Top-5</th>
<th>Top-1%</th>
<th>Hit Rate</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>41.1</strong></td>
<td><strong>65.8</strong></td>
<td><strong>98.4</strong></td>
<td><strong>44.7</strong></td>
<td><strong>11.0</strong></td>
<td><strong>23.6</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Top-1</th>
<th>Top-5</th>
<th>Top-1%</th>
<th>Hit Rate</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>11.0</strong></td>
<td><strong>23.6</strong></td>
<td><strong>80.2</strong></td>
<td><strong>11.6</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparison with State-of-the-art

• Localization in terms of meters
The Effect of Offset Prediction

- Localization in terms of meters

Figure 6. Same-area (left) and cross-area (right) meter-level localization accuracy of different offset prediction methods.
Noisy GPS Refinement

- Retrieval in a searching scope

| Search Scope | Same-Area | | Cross-Area | | |
|--------------|-----------|-----------------|-----------|-----------------|
| | Top-1 | Top-5 | Top-1 | Top-5 |
| All | 41.1 | 65.8 | 11.0 | 23.6 |
| 1000 m | 49.2 | 76.7 | 19.9 | 41.5 |
| 500 m | 54.1 | 82.6 | 26.4 | 53.3 |
| 200 m | 60.9 | 90.6 | 37.7 | 72.0 |
Noisy GPS Refinement

Same-area

Cross-area

Accuracy (%) vs. Threshold (m)

- Original
- All
- 1000 m
- 500 m
- 200 m
Summary

• We propose a new benchmark for cross-view image geo-localization beyond one-to-one retrieval, which is a more realistic setting for real-world applications.

• The proposed method significantly improves 10-meter-level accuracy:
  11.4% → 25.5% for same-area evaluation
  2.8%  →  6.2%  for cross-area evaluation

• We validate the potential of the proposed framework for noisy GPS refinement.
Website:

https://github.com/Jeff-Zilence/VIGOR
Thank you