Object Detection on Street View Images: from Panoramas to Geotags

Vladimir A. Krylov

in collaboration with Eamonn Kenny (TCD), Rozenn Dahyot (TCD)
Motivation. Billions of images (by Google, Bing, Mapillary) covering mlns of kms of road.

~1 mln km coverage

>500 km
Motivation. Billions of images (by Google, Bing, Mapillary) covering mlns of kms of road.

Target. Automatic mapping of stationary recurring objects from Street View.
Object detection. Intro.

➢ Motivation. Billions of images (by Google, Bing, Mapillary) covering mlns of kms of road.

➢ Target. Automatic mapping of stationary recurring objects from Street View.


Mapillary Vistas Dataset
Object detection. Intro.

- **Motivation.** Billions of images (by Google, Bing, Mapillary) covering mlns of kms of road.

- **Target.** Automatic mapping of stationary recurring objects from Street View.

- **State-of-the-art:** Object recognition. **Image geolocation.**

![Query Image](image1.png)  ![Matching Database](image2.png)  
*Lin T. et al., CVPR 2015*

![Example](image3.png)  
*Weyand T. et al., ECCV 2016*
➢ **Motivation.** Billions of images (by Google, Bing, Mapillary) covering mlns of kms of road.

➢ **Target.** Automatic mapping of stationary recurring objects from Street View.

➢ **State-of-the-art:** Object recognition. Image geolocation. **Object geolocation.**

*Wegner, J. et al., CVPR 2016*
➢ Object detection:
Semantic segmentation with Fully Convolutional NNs:

- Introduce extra FP penalty
- Retrain on one or multiple classes of objects: on Mapillary Vistas, Cityscapes

Shelhamer E. et al., IEEE T-PAMI 2017
Spatial scene analysis:

- Stereo-vision, Structure-from-Motion
  - Requires more data, assumptions.

- **Monocular depth estimation**
  - Provides approximate accuracies;
  - Requires segmented objects.

Laina I. et al., 3d Vision 2016
Processing pipeline: geotagging

Strategies to estimate the position of objects from images:

- Depth-based
  - Single view: sensitivity
  - Single view: false positives
  - Low accuracy: up to 7m error

- Triangulation-based
  - High accuracy
  - Multiple views
  - Matching required
We define a **Markov Random Field** (MRF) model over the space of all view-rays intersections:
- label \( z=0 \) if not occupied by object
- label \( z=1 \) if occupied

MRF configuration is characterized by its corresponding energy \( U \). **Optimal = minimum of** \( U \).

Energy terms:
- **Unary term.** *Consistency with depth.*
  \[
  u_1(z) = z \sum_{j=1,2} \| \Delta_j - d_j \|
  \]
- **Pairwise term.** *No occlusions. No spread.*
  \[
  u_2(z) = z \sum_k z_k \| x_k - x_k \|
  \]
- **Ray term.** *Penalize not matched rays.*
  \[
  u_3(z) = (1 - z) \prod_k (1 - z_k)
  \]

Total energy:
\[
U(z) = \sum_{i=1}^{N_z} \left[ \alpha u_1(z_i) + \beta u_2(z_i) + (1 - \alpha - \beta) u_3(z_i) \right]
\]
\( \alpha, \beta \geq 0, \alpha + \beta \leq 1. \)
The geotagging is performed as follows:

- Calculate the space of all intersections;
- Optimize the MRF model;
- Discard non-paired instances;
- Cluster the results. Take intra-cluster averages:
  - *Sparsity assumption.*
Processing pipeline: OVERVIEW

Object detection pipeline:

➢ **DL**: pixel-level **segmentation** to identify objects;

➢ **DL**: monocular **depth** (camera-to-object distance) estimation:
  •  *max distance from camera*: 25m;

➢ **GPS-tagging** based on triangulation and Markov Random field model:
  •  *mild object sparsity assumption* - *1m apart*;

➢ Clustering.
Geotagging of traffic lights in Regent str., London, UK:
- 87 GSV panoramas, 47 out of 50 objects discovered (94% recall)

Map view:

Quantitative performance:

<table>
<thead>
<tr>
<th>Object</th>
<th>#Actual</th>
<th>#Detected</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic Light</td>
<td>50</td>
<td>51</td>
<td>47</td>
<td>4</td>
<td>3</td>
<td>94.0%</td>
<td>92.2%</td>
</tr>
</tbody>
</table>
Results: DEMO

➢ Geotagging of **telegraph poles** over a 2km road, co. Kildare:
  • 170 GSV panoramas, 37 out of 38 objects discovered (97.4% recall)

We gratefully acknowledge financial support and expertise of **eir** in producing these results.
We have developed an image processing pipeline that:

- Is fully automatic;
- The geotagging accuracy comparable with commercial-range GPS-unit;
- Detects and geotags objects at approx. 1.1 GSV panorama per second rate (~3,000 km in 24h on a desktop PC with 2 GPUs);
- Can accommodate custom detection and depth estimation modules.
Thank you!

Contact Us
O'Reilly Building
Trinity College Dublin
Dublin 2
Ireland

adaptcentre.ie