IDENTIFICATION OF BUILDINGS IN STREET IMAGES USING MAP INFORMATION

Masanori Ogawa, Kiyoharu Aizawa
The University of Tokyo

Goal

Building region identification in a street image for embedding geographic information

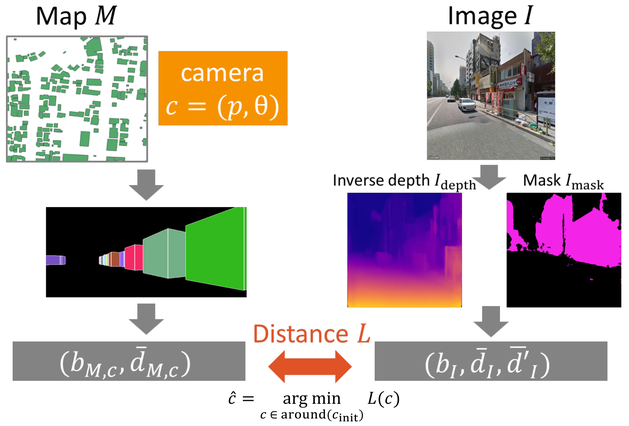
Our Approach

- Given a GPS-based camera pose and a 2D map with building outlines
- Improve the camera pose by using a depth map estimated from the image and the depth information taken from the map
- Project buildings on the map onto the image using the improved camera pose

Related Works

Camera pose refinement/estimation methods using 2D map (Cham et al. [1], Chu et al. [2], Yuan et al. [3]) are based on vertical lines of building corners in the image.

Proposed Method



1. **Building depth calculation from a 2D map**
   A horizontal 1D building depth sequence is calculated based on a camera pose.
2. **Depth estimation of an image**
   Semantic segmentation is used for building mask creation. A horizontal 1D building depth sequence is obtained by using the mask and the estimated depth.
3. **Camera pose refinement**
   The distance between depth sequences are calculated while ignoring where buildings may be covered in the image. The camera pose is searched to minimize the distance.
4. **Building map projection**
   Buildings on the 2D map are projected and masked with the above building mask.

Experiment

Target Images: 80 images of 3 areas (Oshima, Bunkyo, Shinjuku in Japan) via GSV API
2D map: GEOSPACE of NTT Space Information Inc.
Used methods: Monodepth [4] for depth estimation, DeepLab v3+ [5] for building mask
Camera pose candidates:

- $p = p_0 + (x, y)$ $\quad x, y \in [-3\text{m}, 3\text{m}]$ per 0.1m
- Focusing on positions that are noisy in the measurement

Ground truth (GT) preparation:

- GT camera poses are annotated so that the projection results match the images.
- GT building identifier maps are obtained by masking the projection results.

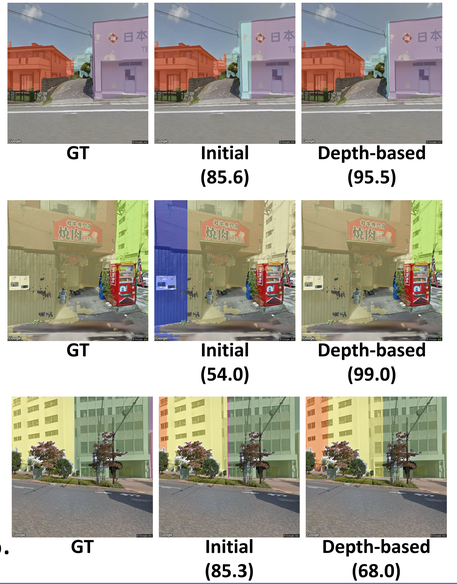
Metrics: Accuracy of building identifiers in the building mask (%)

Evaluation

Method	Oshima	Bunkyo	Shinjuku	All
Initial (GSV)	77.9	76.9	79.9	78.5
Depth-based (Ours)	80.7	81.9	84.5	83.0
Corner-based (modified [3])	79.6	73.5	75.9	75.4
Combined (Ours)	83.0	80.0	86.7	83.5

*Combined: using corner and depth

The building identification accuracy was improved from the results of the initial GSV camera poses in all areas by approximately 5–7%.



[1] T. Cham et al. "Estimating camera pose from a single urban ground-view omnidirectional image and a 2d building outline map," in CVPR. IEEE, 2010.
[2] H. Chu et al. "Gps refinement and camera orientation estimation from a single image and a 2d map," in CVPR Workshops. IEEE, 2014.
[3] J. Yuan and A. M Cheriyadat "Combining maps and street level images for building height and facade estimation," in SIGSPATIAL Workshop. ACM, 2016.
[4] C. Godard et al. "Unsupervised monocular depth estimation with left-right consistency," in CVPR. IEEE, 2017.
[5] L. Chen et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation," in ECCV. Springer, 2018.