Learning to Align Semantic Segmentation and 2.5D Maps for Geolocalization

Anil Armagan, Martin Hirzer, Peter M. Roth, Vincent Lepetit
Institute of Computer Graphics and Vision, Graz University of Technology
{armagan, hirzer, roth, lepetit}@icg.tugraz.at

Contribution
We present an efficient method for geolocalization in urban environments starting from a coarse pose estimate provided by GPS and compass information and using a simple untextured 3D model of the surrounding buildings. We train two deep networks to predict the best direction to improve the pose estimate, given a semantic segmentation of the input image and a rendering of the buildings from this estimate. We then iteratively apply these networks until converging to a good pose.

2.5D Maps
• Light weight 3D models
• Ground plane locations of buildings' corners
• Height of each building
• No texture or color
• Available on OpenStreetMap

Relation Between Input Image and 2.5D Map
• How to construct the relation between the image and the model?
• We use FCN [1] to semantically segment the input image into related classes: façade, vertical edge, horizontal edge and background

System Overview
• Both networks are iterated until their predictions are “don’t move” and “don’t rotate” at the same time.

Learning to Predict a Direction for Pose Update
• Translation network: 8 directions or “don’t move”
• Orientation network: 2 directions or “don’t rotate”
• Example of a pose update by the Translation network:

 Pose Log-Likelihood \( \mathcal{L}(\text{pose}) = \sum x \log P_{\text{class}}(\text{render}(\text{pose}), x) (x) \)

Results
• We tested our approach on 40 test images and evaluated the position and orientation errors of \( \hat{p} \) and the pose found by our method w.r.t. the ground truth pose. Our method decreases the mean orientation error of \( \hat{p} \) from 11.3° to 3.2° and the mean position error from 13.4m to 3.1m.

References

Acknowledgment
This work was funded by the Christian Doppler Laboratory for Semantic 3D Computer Vision.