Sensor Fusion for Semantic Segmentation of Urban Scenes
Richard Zhang\textsuperscript{1} Stefan A. Candra\textsuperscript{1} Kai Vetter\textsuperscript{1,2} Avideh Zakhor\textsuperscript{1}
\textsuperscript{1}Department of Electrical Engineering and Computer Science, UC Berkeley
\textsuperscript{2}Department of Nuclear Engineering, UC Berkeley

Introduction

Goal: effectively fuse information from multiple modalities to obtain semantic information

Contributions:
- information from multiple scales considered
- late fusion used to maximally leverage training data
- validated on KITTI data \cite{Geiger2013} with augmented labels; performance improvements obtained over state-of-the-art method \cite{Cadena2014}

Multi-Level Segmentation
- Multiple segmentations to consider cues from varying scales of information in classification
- Image: hierarchical segmentation \cite{Arbelaez2014} extracted
- Point cloud: 0.5 m supervoxels and connected component segmentation

Feature extraction
- Inference performed on small-scale segments
- Small-scale segments associated with large-scale segments
- Feature vectors of small-scale segments augmented with associated large-scale segment

Features Extracted

- Point cloud supervoxel features
- Image superpixel features

Classification & Late-Fusion
- Random Forest (RF) classifier used for each modality separately
  - PMFs of unimodal classifications
  - PMFs serve as compact and descriptive mid-level features
  - Post-processing pairwise CRF to provide spatial smoothing

Late-fusion Results
- Fusion improves performance for overlapping regions:
  - Img: 82.1% pc only, 87.7% fused
  - PC: 84.9% img only, 99.9% fused
  - Examples: sidewalk more likely to be classified correctly vs road only after fusion

Conclusions
- Dataset: 252 images (140 training, 112 testing) from 8 sequences
- Multiscale information provides strong cues for classifier
- late fusion greatly boosts performance
- Outperforms current state-of-the-art \cite{Cadena2014}
- Stuff classes well discriminated

Path Forward
- Add 2D+3D object detectors to increase performance on things
- Enforce consistency across temporal and 3D spatial dims
- Extend algorithm to additional modalities, e.g. infrared and hyperspectral, and validate
- Integrate with reconstruction algorithms

Qualitative Results

References
\cite{Geiger2013, Arbelaez2014, Cadena2014}