Report for Scientific Initiative 2017
ISPRS Benchmark on UAV Semantic Video Segmentation

Principle Investigators

Dr. Michael Ying Yang
Assistant Professor, ITC, University of Twente, the Netherlands
michael.yang@utwente.nl
Co-Chair of ISPRS WG II/5: Dynamic Scene Analysis

Dr. Alper Yilmaz
Professor, PCV Lab, The Ohio State University, USA
yilmaz.15@osu.edu
Chair of ISPRS WG II/5: Dynamic Scene Analysis

Summary

Visual understanding of complex urban scenes is an enabling factor for a wide range of applications. ISPRS Benchmark on UAV Semantic Video Segmentation aims to promote and advance the video segmentation task of VHR UAV sequences. The computer vision community has relied on several centralized benchmarks for performance evaluation of numerous tasks. Such benchmarks have proved to be extremely helpful to advance the state-of-the-art in the respective research fields. There has been rather limited benchmark effort on the ISPRS community. With this benchmark we would like to pave the way for a unified framework towards meaningful quantification of semantic segmentation from UAV imagery and videos.
Currently, 10 videos have been captured by DJI Phantoms. In total, 75,000 frames have been acquired in Germany and China. We select eight classes, i.e. building, road, tree, low vegetation, moving car, static car, human and background. Annotation and quality control require more than 2 hours for one frame. So far, two videos has been labeled. The rest of the videos will be labeled later in this year.
We will provide a comprehensive benchmark suite later this year by: (i) creating the largest dataset of UAV scenes with high-quality annotations; (ii) developing a sound evaluation methodology for pixel-level semantic labeling; (iii) setting up a corresponding challenge. It is expected that this envisaged benchmark will enhance the visibility of ISPRS research and events, attracting Computer Vision researchers joining ISPRS communities.

Activities from March 2017 till February 2018

The major labor work went into the creation of data for the UAV Semantic Video Segmentation benchmark. In particular the following steps have been performed:

1. UAV data capture

Our UAV video dataset is captured in 4K mode, and each frame has resolution 4096x2160. The camera has 45 degrees angle to the horizon plane, giving access to capturing a large amount of objects. Currently, 10 videos have been captured by DJI Phantoms. In total, 75,000 frames have been acquired in Germany and China.

2. Annotation

Since there is no good publically available annotation tool for video segmentation, our main effort was to create video labeling software based on C++ with Matlab interface. We focus on seven relevant urban classes, i.e. *building, road, tree, low vegetation, moving car, static car, human (plus background/clutter)*. Class definitions are as follows:

- Building: living houses, garages, skyscrapers and so on.
- Road: only public road that cars can run on. Small road connecting main road and garage is not included.
- Tree: tall trees that have clear canopy and main branches.
- Low vegetation: grass, bushes and shrubs.
- Moving car: car that is moving.
- Static car: car that is not moving.
- Human: pedestrians or cyclists.
- Background: class not belonging to any class above.

Figure 1 shows video navigation interface. The software provides basic video playing and video frame labeling.

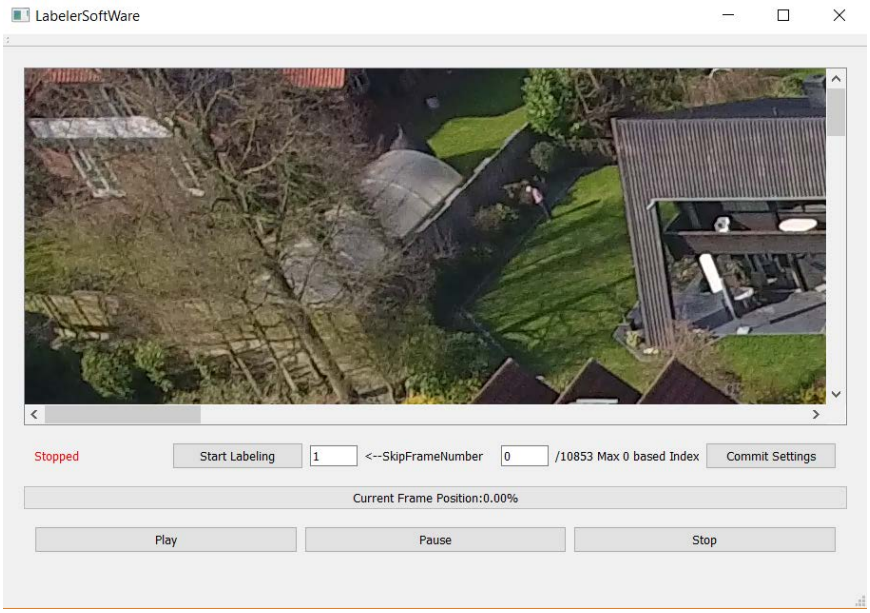

Figure 1: Video navigation interface.

Figure 2 shows our software interface. In order to annotate in different accuracy and pace, we have developed different labeling strategies in our software: pixel level labeling, super pixel level labeling and polygon mode labeling.

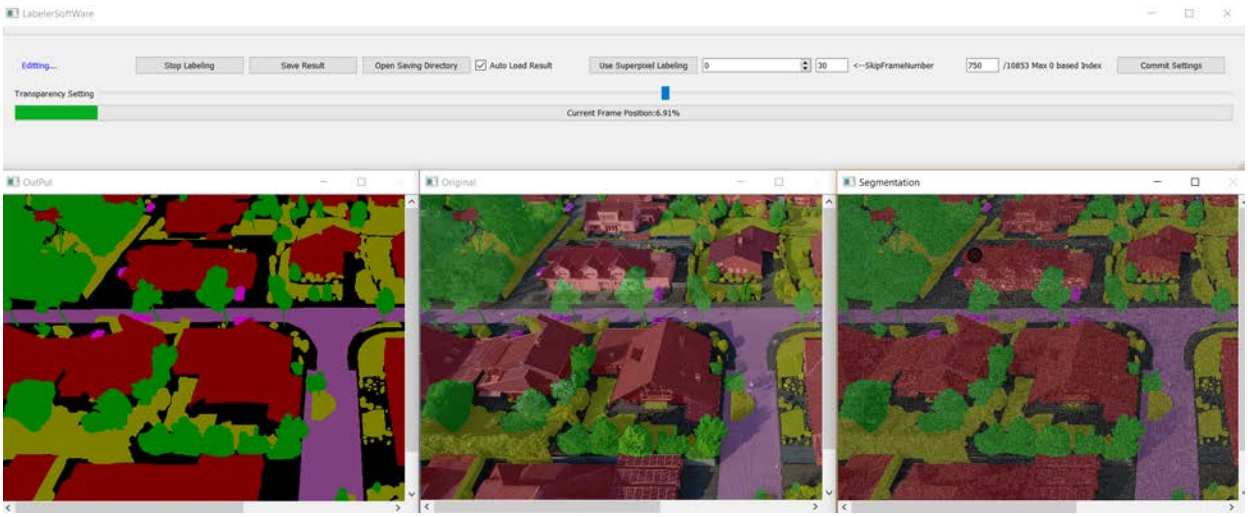

Figure 2: Software interface.

Example of a UAV image and the corresponding label result are shown in Figure 3. Since the data is of very high resolution and due to the complexity, the labeling needed much more attention than estimated. Annotation and quality control require more than 2 hours for each frame. The rest of the videos will be labeled later in this year.

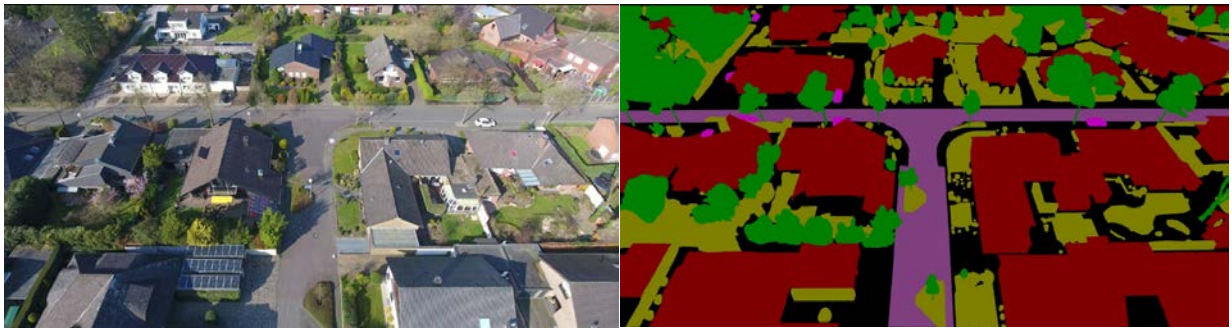

Figure 3: Example of a UAV image and the corresponding label results.

Currently, three videos has been fully labeled for every 150 frames, corresponding to about 5 second interval. Example of a UAV video and the corresponding label results are shown in Figure 4.

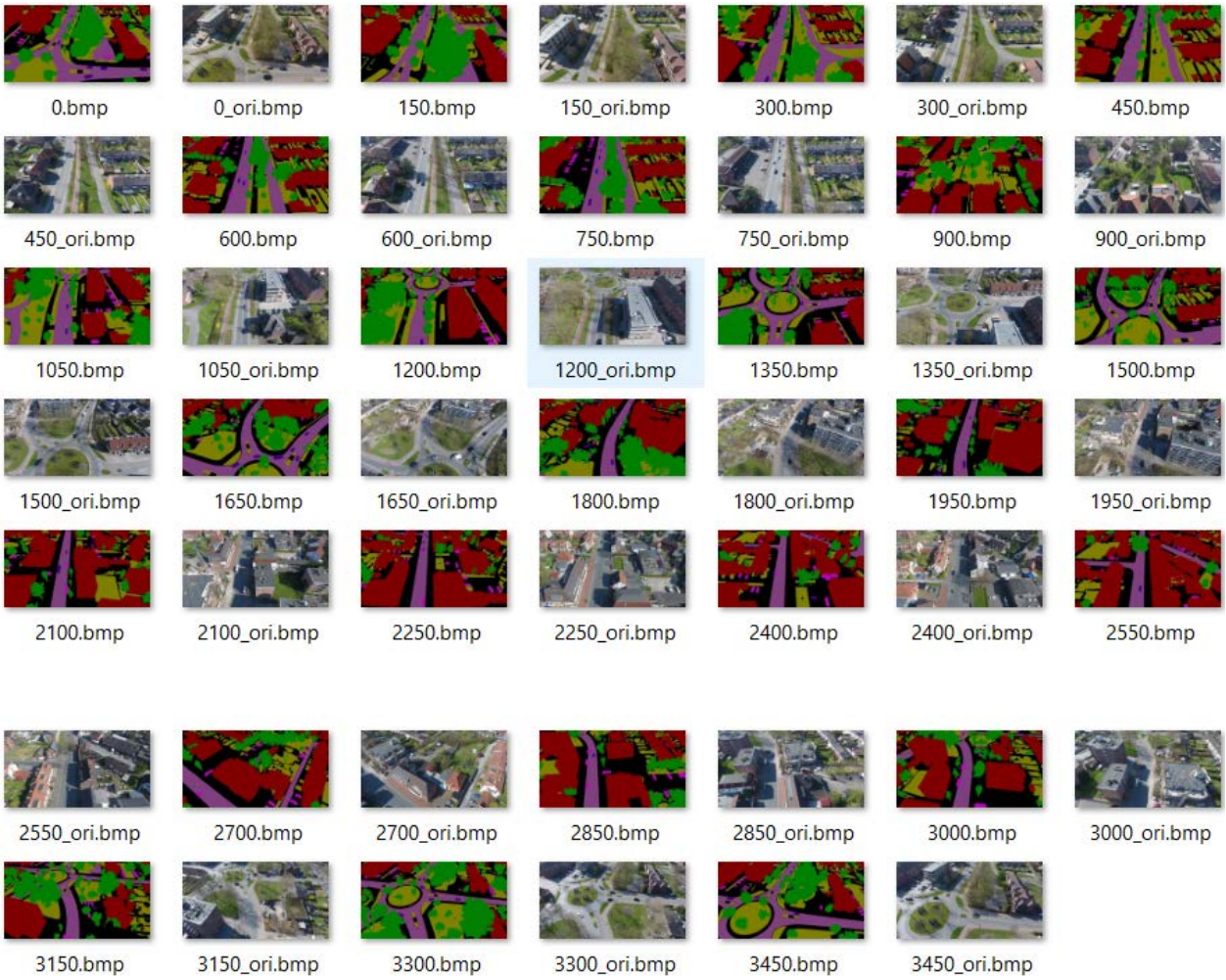

Figure 4: Image label pairs for one example video.

Outlook

Although the financial support by ISPRS ends this year we plan to continue labeling of the rest videos. We plan to provide a comprehensive benchmark suite later this year by: (i) creating the largest dataset of UAV scenes with high-quality annotations; (ii) developing a sound evaluation methodology for pixel-level semantic labeling; (iii) setting up a corresponding challenge.

Justification of money spent in 2017

Total CHF 7,000 were available for 2017. In 2017, 10 videos have been captured by DJI Phantoms. In total, 75,000 frames have been acquired in Germany and China. The labor cost is CHF 1,500. Since the data is of very high resolution and the complexity, the processing but also the labeling needed much more attention than estimated. Annotation and quality control require more than 2 hours for one frame. The total costs for student assistants for ground truth data labeling in 2017 is CHF 7,500. The difference of CHF 2,000 has been covered by the budget of the ITC Faculty of University of Twente.