Technical Report

Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)

Markus Gerke, ITC, University of Twente
m.gerke@utwente.nl

Introduction

This report describes experiments conducted using the multi-class image classification framework implemented in the stair vision library (***SVL***, (Gould et al., 2008)) in the context of the ISPRS 2D semantic labeling benchmark. The motivation was to get results from a well-established and public available software (Gould, 2014), as a kind of baseline. Besides the use of features implemented in the SVL which makes use of three channel images, assuming RGB, we also included features derived from the height model and the NDVI which is specific here, because the benchmark dataset provides surface models and CIR images. Another point of interest concerned the impact the segmentation had on the overall result. To this end a pre-study was performed where different parameters for the graph-based segmentation method introduced by Felzenszwalb and Huttelocher (2004) have been tested, in addition we only applied a simple chessboard segmentation. Other experiments focused on the question whether the conditional random field classification approach helps to enhance the overall performance.
The official evaluation of all experiments described here is available at
http://www2.isprs.org/vaihingen-2d-semantic-labeling-contest.html (SVL_1 to SVL_6). The normalized height models are available through the ReseachGate profile of the author (http://www.researchgate.net/profile/Markus_Gerke)

Workflow

The stair vision library offers a set of tools to perform the following actions being relevant here, see (Gould et al., 2008) and the svlBook available at (Gould, 2014).

- Extraction of image-based features (like Haarfeatures, Textons, Spin, Rift) in 3-channel images
- Training of an Adaboost-based classifier using (a subset) of the images and ground truth labeling and prediction of unseen images
- Training of a Condition Random Field model using the classification result from the Adaboost-step, prediction of unseen images

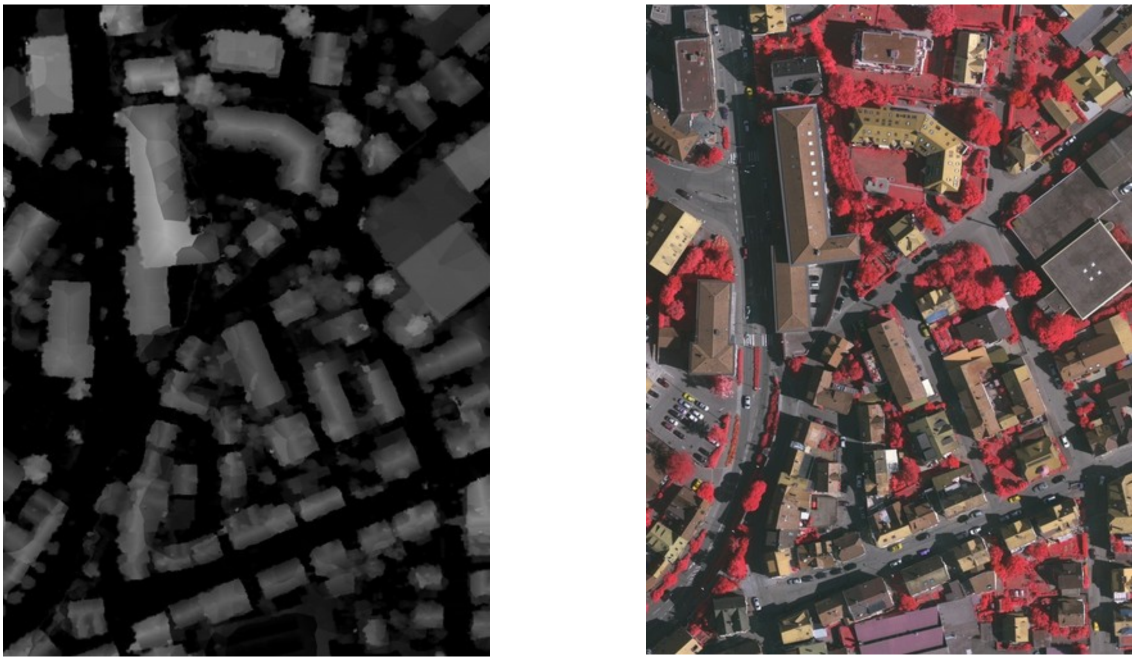
The entities to be classified ("superpixels" or segments) need to be provided through other means, there is no method available in SVL to do that. In our experiments we used the graph-based approach by Felzenszwalb and Huttelocher (2004), who also publish the source code of their method. In addition we also test a simple chessboard segmentation.

In order to better exploit the image information provided in the context of the benchmark we computed the following features in addition to the standard SVL-features:

- NDVI: the normalized digital vegetation index, computed from the first (IR) and second channel (R) of the CIR true ortho photo (TOP).
- saturation: some experiments revealed that the saturation of the CIR image is helpful to further support the separation of vegetation (well saturated) and impervious surfaces.
- Normalized height: the digital surface model (DSM) provided is classified into ground and off-ground pixels using the lastools-toolbox (http://rapidlasso.com/lastools/), which uses an improvement of the method by Axelsson (2000). For all off-ground pixels the closest ground point is assumed to be the relevant low point and thus through reduction of the height of the off-ground point by the assigned ground point the so-called normalized height is computed. In this final representation

the influence of varying ground heights into the classification gets removed. This method does not compute correct normalized heights since local constraints like horizontal ridge lines are not considered, hence, when ground heights around one building are varying, the heights are not correct, see large L-shaped building in Fig. 1. The influence of those variations on the classification result has not been analyzed.

Attention! The filtering was performed using the lasground tool in a batch process. There are still partly quite some large errors, e.g. in some tiles (like in tile 31) large industrial halls have been labeled as ground and are thus not included in the normalized DSM. Those errors, are, however, not corrected because we wanted to show the performance of a fully automatic workflow.



Normalized DSM

CIR true ortho photo

Figure 1: Example normalized DSM and corresponding true ortho photo, area 1

Experiments

Assessment of image-based segmentation

The optimal segmentation should result in a balance of over- and under-segmentation of the scene. If the over-segmentation is strong, objects might get too fragmented. In turn, under-segmentation might result in merged objects.

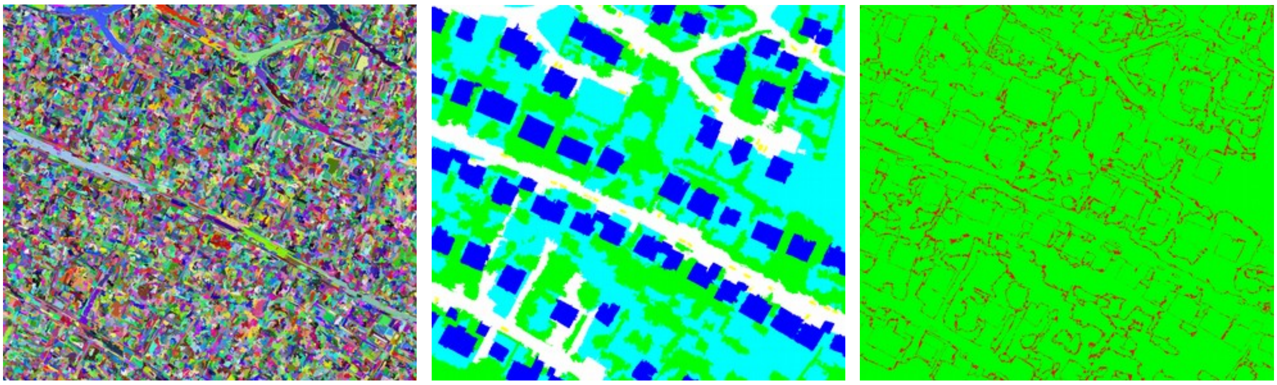
The graph-based segmentation approach (Felzenszwalb and Huttenlocher, 2004) efficiently groups pixels similar to the perceptual appearance within a local neighborhood. The algorithm balances intra-segment and inter-segment heterogeneity differences, and thus can be easily adjusted to ignore certain pixel-to-pixel differences if in a local neighborhood they are part of a homogenous pattern. Three parameters need to be defined, see also Felzenszwalb and Huttelocher (2004):

- sigma: prior to the actual segmentation the image gets smoothed using a Gaussian with sigma as the standard deviation. The larger sigma, the larger the segments since small variations are smoothed out. If sigma is too large the risk of under-segmentation increases.
- k: the segmentation algorithm iteratively merges adjacent pixel-components (segments) if a similarity threshold function is satisfied, formulated as an edge weight. The threshold is depending on k, where larger values for k will cause preference for larger segments.
- min: minimum segment size (number of pixels), applied in a post-processing step.

In order to evaluate the segmentation the following workflow has been implemented:

- for one tile (here area 13) do the segmentation with some selected parameters. The selected tile also has ground truth labeling available.
- for each segment assign a label according to a majority vote. This results in an artificial labeling image.
- Assess the artificial labeling image on a pixel basis, i.e. compute the confusion matrix. One by-product of this evaluation is a red/green image where wrongly assigned pixels got colored red.
- The red/green image and the segmentation image enable an assessment of over- and under-segmentation.

Several parameter combinations have been systematically tested. Figure 2 shows the segmentation, artificial label image and red/green for the finally chosen segmentation parameters (sigma=0.1, k=150, min=20).



Segmentation

Reference labels assigned to segments

Pixel-based evaluation

Figure 2: Segmentation in TOP, area 13, assignment of labels and pixel-based evaluation.

Using those selected parameters leads to an overall "classification" accuracy of 93.4%, the worst result is obtained for the class "car" which is around 83%. It is important to note that the final classification using such a segmentation cannot be better than this artificial labeling using the reference labels.

Figure 3 reveals that especially the transition from low vegetation to trees and from impervious surface to cars/buildings does not get sharply delineated.

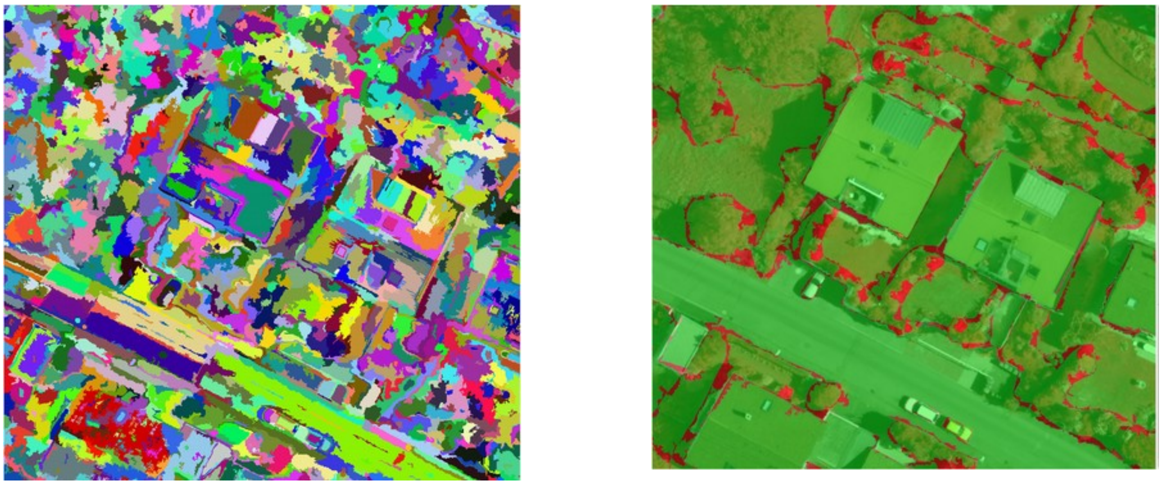
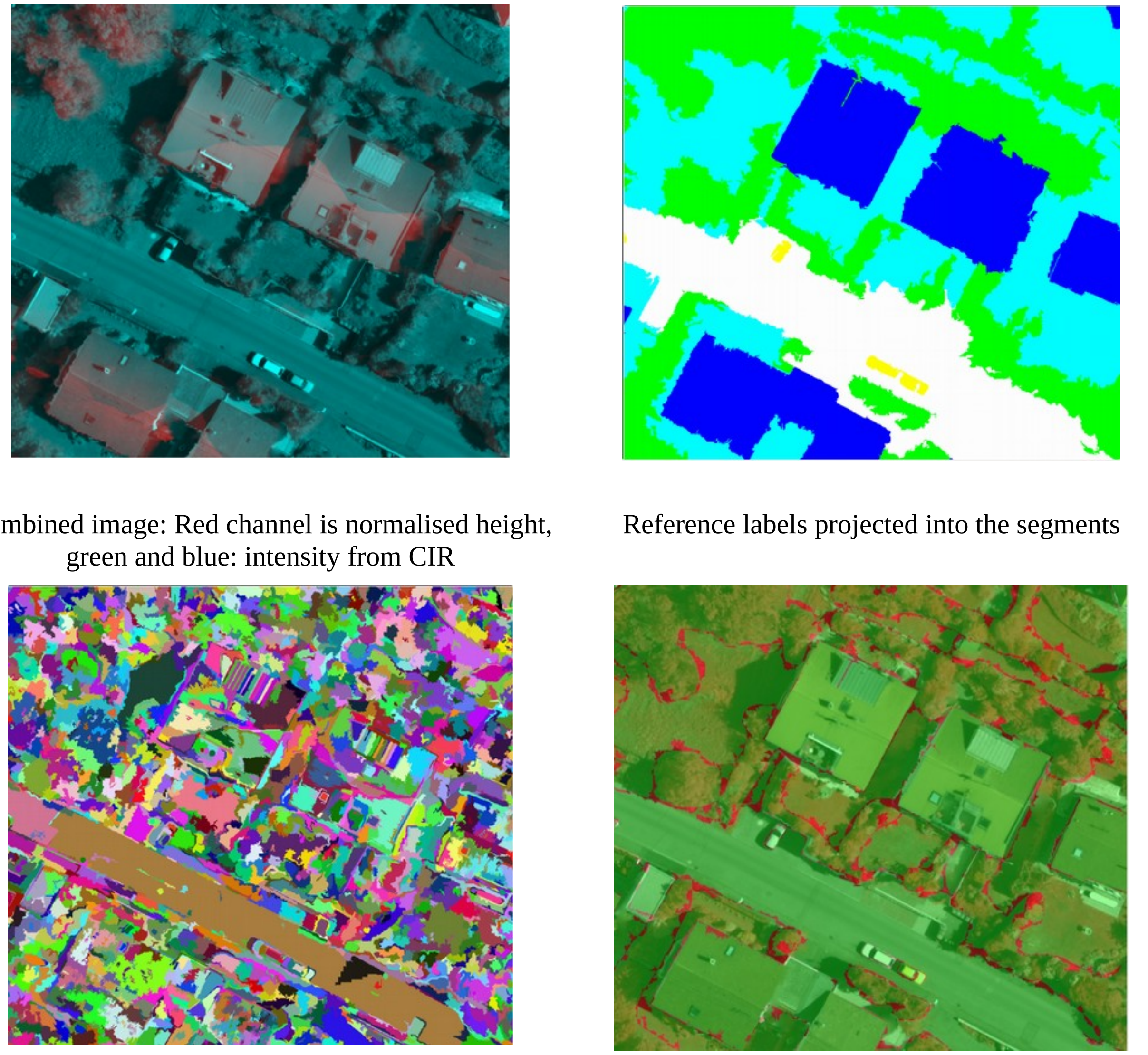


Figure 3: Zoom in to segmentation and overlay of red/green image with CIR TOP

In order to test if the segmentation can be better adjusted to those transitions from ground to off-ground objects, the normalized height model got combined with the CIR image and this merged image was used for the segmentation. The image used for the segmentation is composed as follows: red channel: normalized height model (contrast-stretched), green and blue channel: intensity of original CIR image. See Figure 4, top left image for an example.



Combined image: Red channel is normalised height, green and blue: intensity from CIR

Reference labels projected into the segments

Segmentation in the combined image

Red/green evaluation overlaied with original CIR

Figure 4: Segmentation in the combined normalized height/CIR image

The exemplary results of the segmentation in the combined image are shown in Figure 4. Compared to the former segmentation where only the CIR is employed, the delineation of objects is less smooth here because of the impact from the height model (see for instance building outline in upper right image in Figure 4). In addition, in some areas the height jumps are not maintained, leading to an under-segmentation. However, the accuracy of labeling is better by around 1%, compared to the initial version. In the classification experiments (SVL_4) these segments will be used and compared to the classification in the initial segmentation.

Chessboard segmentation

Another approach is to skip pre-segmentation and directly classify each individual pixel. Compared to an image-based pre-segmentation this has the advantage that an under-segmentation as observed above cannot affect the classification performance. The downside is, however, that because many more instances are available for training the classifier, the memory and processing time consumption will increase considerably. Given the size of the images a per-pixel classification is not possible with the SVL. In order to test such an alternative strategy we have performed a chessboard segmentation instead, that is all pixels in a *nxn* neighboorhood form one entity. We found that for this application and given our IT infrastructure a cluster size of 5x5 (i.e. 45cmx45cm) to be a good compromise: object outlines are still sufficiently represented, i.e. not too coarsely delineated, and the computation time for the adaboost training is not excessive (though still *1 day* for this data, compared to *2hours* for the image-based presegment case). For the classification all features (compare SVL_1) have been used. The results are in SVL_5 (only adaboost) and SVL_6 (with crf). See analysis below.

Classification experiments

Based on the selected parameters all images have been segmented and the features have been computed, both the original SVL-features, and in addition the CIR-image features NDVI and saturation, and normalized heights, derived from the DSM. All tiles where ground truth is provided were used to train the classifier, while all the other tiles were classified and evaluated using the official evaluation procedure: http://www2.isprs.org/semantic-labeling.html#Vaihingen2D_label_eval
Three main experiments have been conducted using the segmentation in the CIR image only:
SVL_1: SVL-features, including NDVI, saturation and normalized height, boosting and CRF classification
SVL_2: SVL-features only, boosting and CRF classification
SVL_3: SVL-features, including NDVI, saturation and normalized height, only boosting

One experiment has been done using the segmentation of the combined height/CIR image (Fig. 4):
SVL_4: SVL-features, including NDVI, saturation and normalized height, boosting and CRF classification

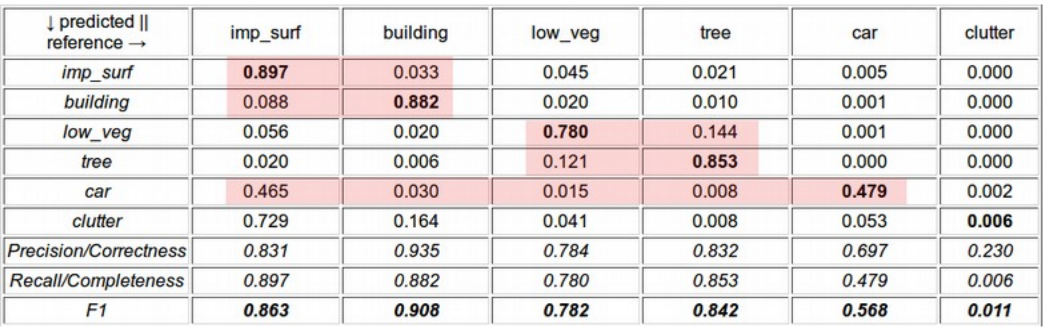
Two last experiments concern the use of a chessboard segmentation with 5x5 pixel raster size:
SVL_5: SVL-features, including NDVI, saturation and normalized height, boosting only classification
SVL_6: SVL-features, including NDVI, saturation and normalized height, boosting and CRF classification

This means that *SVL_1 vs. SVL_2* is to compare the impact of additional features derived from CIR and height, while *SVL_1 vs. SVL_3* indicates the impact the CRF has on the classification result, i.e.: is the conditioning on neighborhood relations helpful in this context?
SVL_1 vs. SVL_4 analyses the impact the height model has on the segmentation; all the features are identical for SVL_1 and _4.

SVL_5 and *SVL_6* use the same features as *SVL_1*, but the image-based segmentation is skipped and the classification is done in the chessboard-segmented images. SVL_5 (no CRF) vs. SVL_6 (with CRF) is to assess the impact of CRF has in this case.
For the assessment of classification results we use the reference set without object boundaries (right column in the details webpages). The classes are labeled as follows:

⬜	Impervious surface	🟦	Low vegetation	🟨	Car
🟪	Building	🟩	Tree	🟥	Clutter

CRF-based classification (SVL_1, _2), segmentation in CIR image only



↓ predicted || reference →	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.897**	0.033	0.045	0.021	0.005	0.000
building	0.088	**0.882**	0.020	0.010	0.001	0.000
low_veg	0.056	0.020	**0.780**	0.144	0.001	0.000
tree	0.020	0.006	0.121	**0.853**	0.000	0.000
car	0.465	0.030	0.015	0.008	**0.479**	0.002
clutter	0.729	0.164	0.041	0.008	0.053	**0.006**
Precision/Correctness	*0.831*	*0.935*	*0.784*	*0.832*	*0.697*	*0.230*
Recall/Completeness	*0.897*	*0.882*	*0.780*	*0.853*	*0.479*	*0.006*
F1	***0.863***	***0.908***	***0.782***	***0.842***	***0.568***	***0.011***

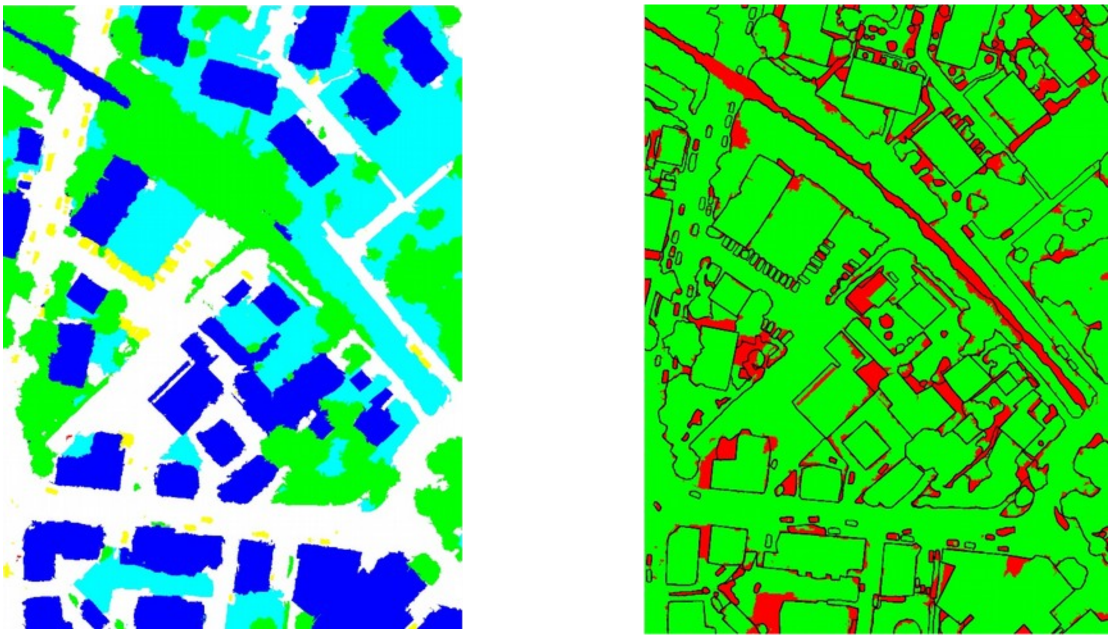


Figure 5: Results from SVL_1: SVL-features, including NDVI, saturation and normalized height, boosting and CRF classification, upper: overall confusion matrix and classification result and red/green evaluation for tile 8

The upper table in Figure 5 shows the overall results for SVL_1 (all SVL features plus NDVI, saturation, normalized height), using the CRF classification method. The completeness for impervious surfaces and buildings is almost 90% while the vegetation classes are 78% (low vegetation) and 85% (trees). Cars are much worse, only 48% completeness. Although the correctness of cars is better (around 70%) the confusion of cars with impervious surfaces is obvious: almost 50% of actual cars-pixel have been classified as impervious surfaces. Approximately 12% of tree pixels have been classified as low vegetation, but as shown in the segmentation experiment those errors are to some extent resulting from a wrong segmentation in the transition area of ground and off-ground areas.

The analysis of the visualization of one label image (in this case tile 8) helps to further interpret the result, see Figure 6 for the original CIR image of area 8. One obvious confusion concerns impervious surface and low vegetation. Especially if vegetated areas are in shadow they are often classified as impervious surfaces. The same observation was made in the context of the object detection benchmark, see Rottensteiner et al. (2014).

In SVL_2 the NDVI, saturation and height features have not been used, else this setup was identical to SVL_1. Figure 7 shows the results in the same arrangement as for SVL_1. As could be expected we can observe some much larger confusion between above and on-ground features (impervious surface vs. buiding), (low vegetation vs. tree), however, within classes on ground (impervious surface, low vegetation) and above ground (tree, buildings) the confusion is in the same range as for SVL_1. This can be a hint that the height plays a major role, but NDVI/Saturation are not so helpful, compared to the standard SVL features. One explanation is that although the SVL does not compute the NDVI, it anyhow uses the CIR image and thus can separate vegetation from non-vegetation easily. Looking at Figure 7 it becomes obvious that even entire buildings might be missed: the flat roof hall in the center of the tile got classified as impervious surface, probably because the surface appearance and texture is similar to the asphalt.

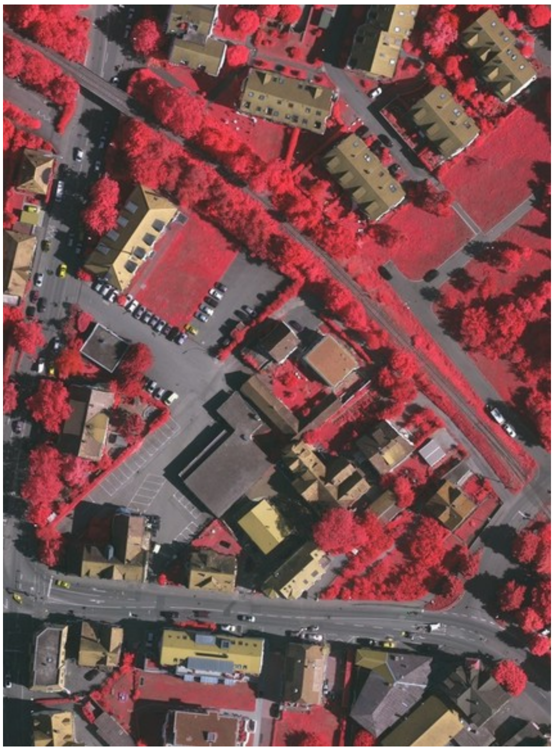
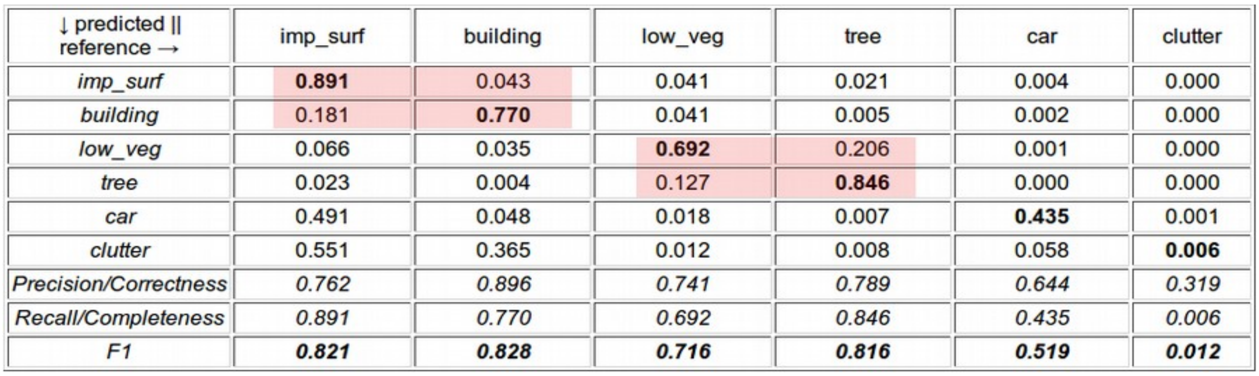


Figure 6: original CIR TOP of area 8



↓ predicted || reference →	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.891**	0.043	0.041	0.021	0.004	0.000
building	0.181	**0.770**	0.041	0.005	0.002	0.000
low_veg	0.066	0.035	**0.692**	0.206	0.001	0.000
tree	0.023	0.004	0.127	**0.846**	0.000	0.000
car	0.491	0.048	0.018	0.007	**0.435**	0.001
clutter	0.551	0.365	0.012	0.008	0.058	**0.006**
Precision/Correctness	*0.762*	*0.896*	*0.741*	*0.789*	*0.644*	*0.319*
Recall/Completeness	*0.891*	*0.770*	*0.692*	*0.846*	*0.435*	*0.006*
F1	***0.821***	***0.828***	***0.716***	***0.816***	***0.519***	***0.012***

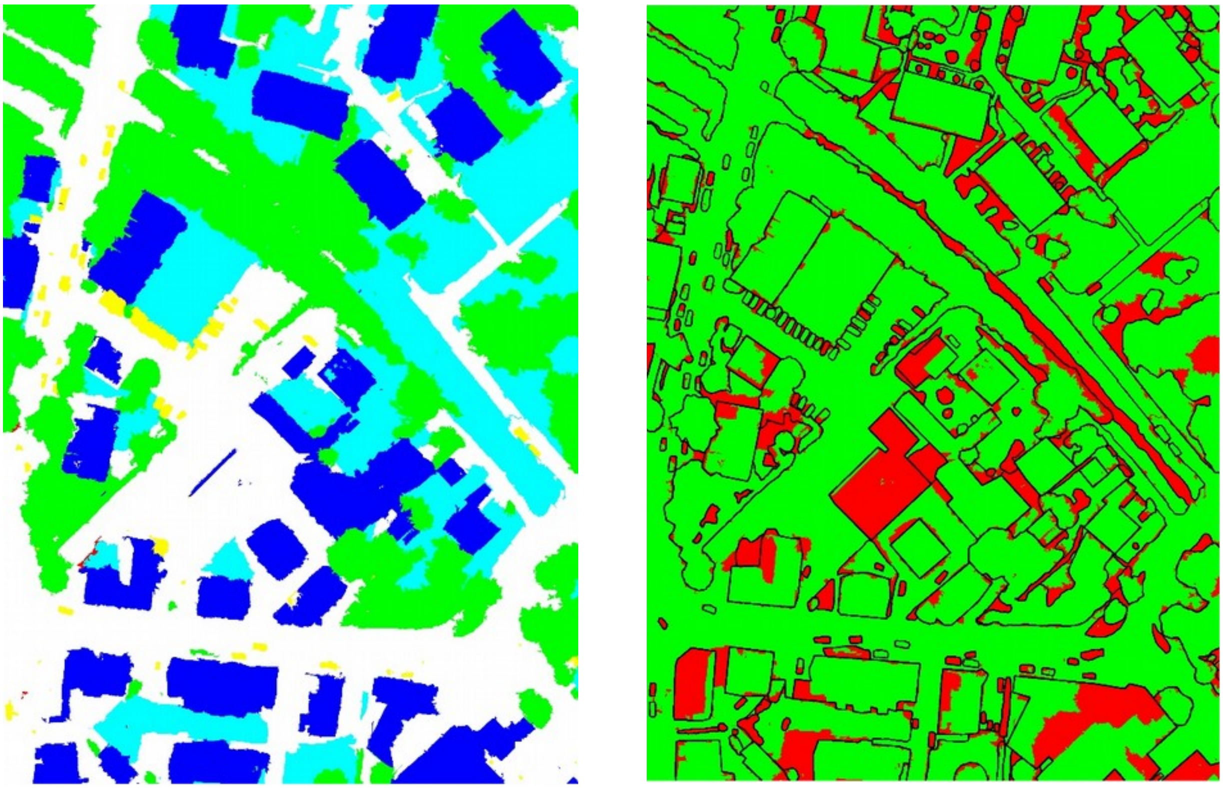


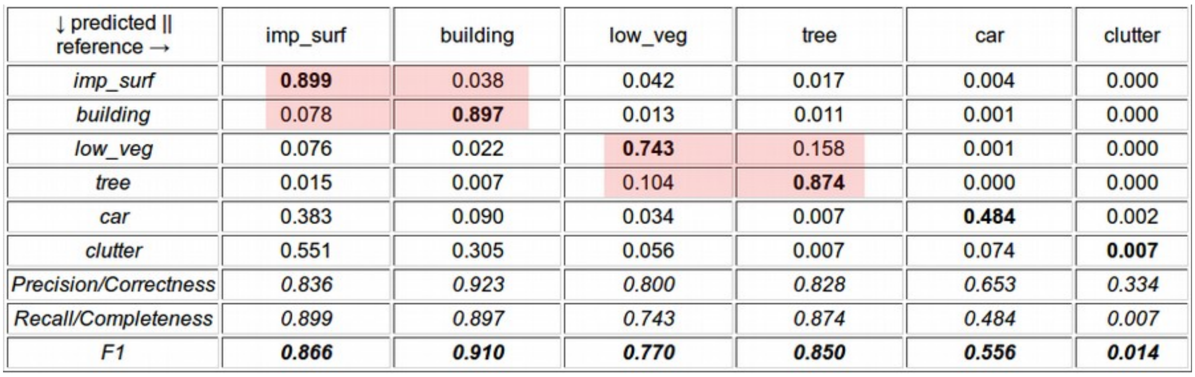
Figure 7: Results from SVL_2: SVL-features only, boosting and CRF classification

Comparing SVL_1 and SVL_2 we can conclude that the (normalized) height is a very important feature, while the computation of the NDVI does not seem to have a positive impact on the overall result.

Whether the height feature is included or not has no large impact on the detection rate of car-pixels. There are possibly four explanations. First, the mean height difference of a car segment compared to ground segments is not significantly large enough to influence the decision during boosting. Second, the ground filter algorithm might label those relatively small discontinuities as ground features and hence some cars might not be represented in the normalized DSM. In addition moving cars (like any non-static objects) are anyhow not included in the dense matching point cloud, i.e. only a subset of all visible cars are represented. Last not least, the segmentation may have under segmented the scene - some cars might have been merged with the background.

Boosting-only- classification, segmentation in CIR image only (SVL_3)

As far as the result from SVL_3, i.e. same features as SVL_1 but skipping the final CRF step and just using Adaboost, is concernd, we can make some interesting observations, as well. According to the evaluation measures this result is slightly better than the one with CRF enabled (SVL_1), this means the smoothing effect induced by CRF seems to have a negative impact. However, looking at the example label image (Fig. 8) reveals that without CRF many objects show some typical speckle effect and thus for the practical use the CRF result might be better suited.



↓ predicted || reference →	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.899**	0.038	0.042	0.017	0.004	0.000
building	0.078	**0.897**	0.013	0.011	0.001	0.000
low_veg	0.076	0.022	**0.743**	0.158	0.001	0.000
tree	0.015	0.007	0.104	**0.874**	0.000	0.000
car	0.383	0.090	0.034	0.007	**0.484**	0.002
clutter	0.551	0.305	0.056	0.007	0.074	**0.007**
Precision/Correctness	*0.836*	*0.923*	*0.800*	*0.828*	*0.653*	*0.334*
Recall/Completeness	*0.899*	*0.897*	*0.743*	*0.874*	*0.484*	*0.007*
F1	***0.866***	***0.910***	***0.770***	***0.850***	***0.556***	***0.014***

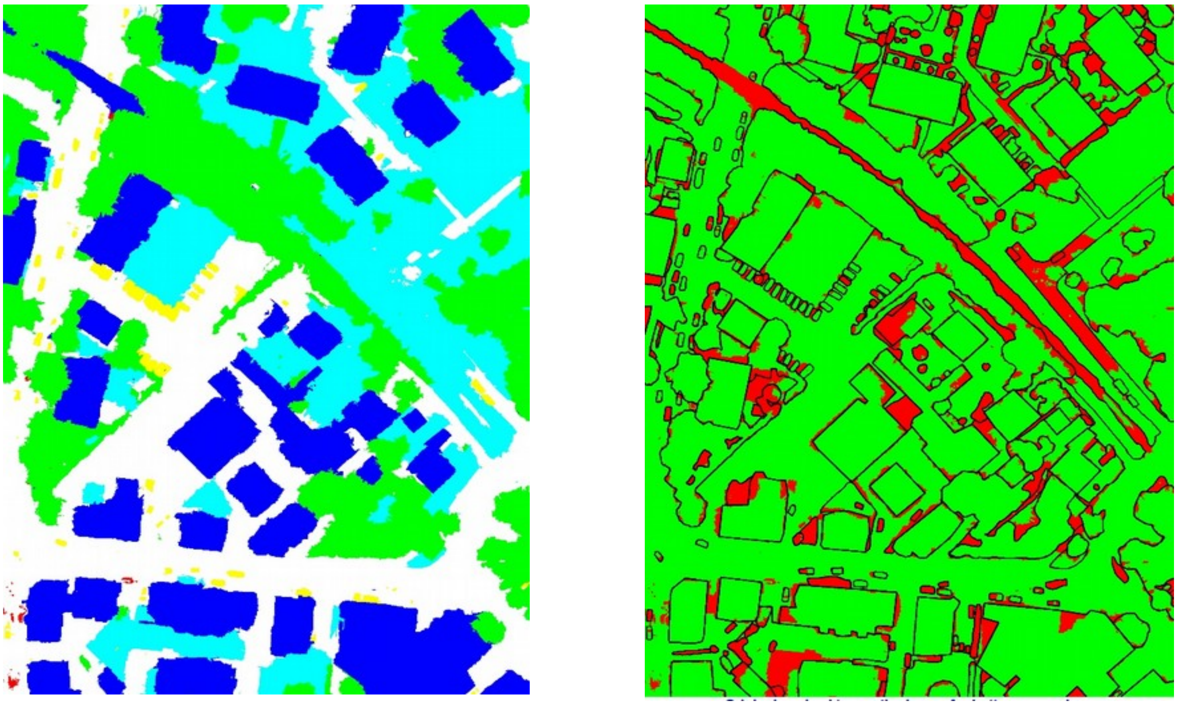
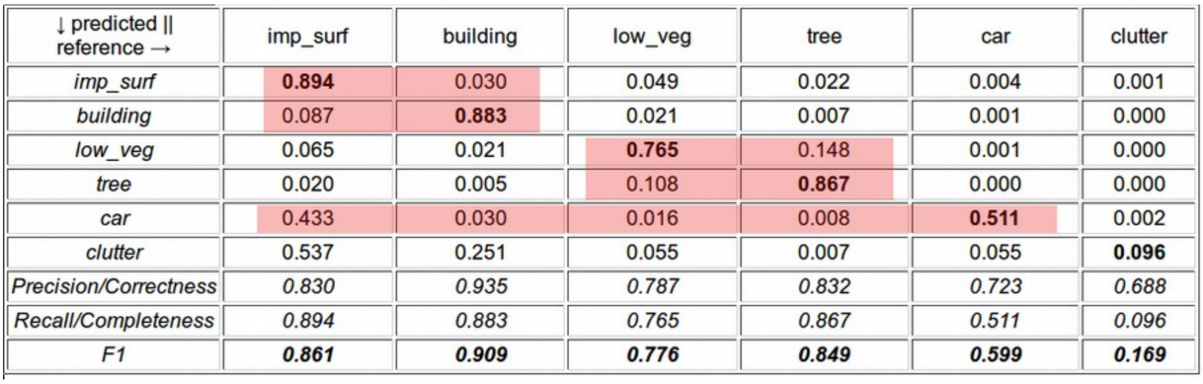


Figure 8: Results from SVL_3: SVL-features, including NDVI, saturation and normalized height, but only boosting classification

CRF-classification using segmentation in height-model/CIR combined image (SVL_4)

As discussed in the experiments on the segmentation algorithm a different segmentation has been performed in images, where the normalized height model was used as red channel, and the CIR image intensity as green/blue channel, refer to Figure 4. The results using this scenario are shown in Fig. 9.



↓ predicted || reference →	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.894**	0.030	0.049	0.022	0.004	0.001
building	0.087	**0.883**	0.021	0.007	0.001	0.000
low_veg	0.065	0.021	**0.765**	0.148	0.001	0.000
tree	0.020	0.005	0.108	**0.867**	0.000	0.000
car	0.433	0.030	0.016	0.008	**0.511**	0.002
clutter	0.537	0.251	0.055	0.007	0.055	**0.096**
Precision/Correctness	*0.830*	*0.935*	*0.787*	*0.832*	*0.723*	*0.688*
Recall/Completeness	*0.894*	*0.883*	*0.765*	*0.867*	*0.511*	*0.096*
F1	***0.861***	***0.909***	***0.776***	***0.849***	***0.599***	***0.169***

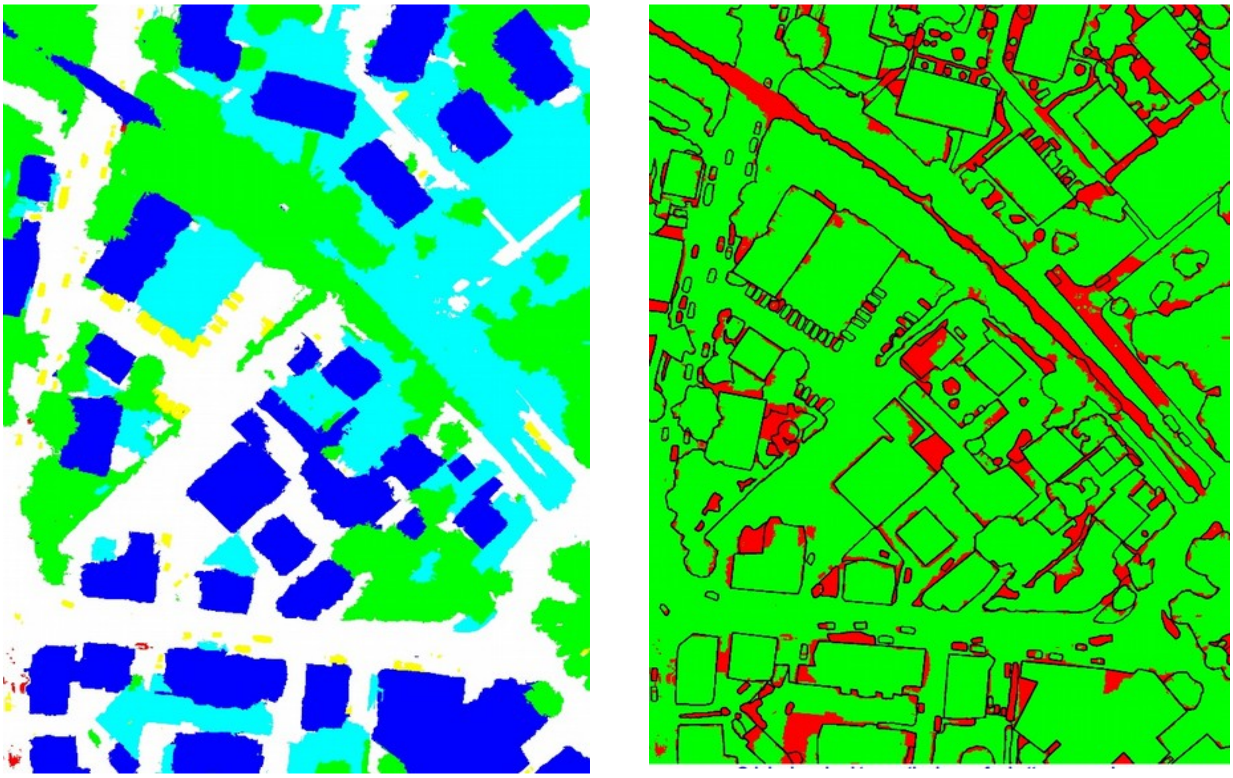


Figure 9: Results of SVL_4: same features and classification method as in SVL_1, but segments from height/CIR-fused image

The overall results are comparable to SVL_1, and also the confusion between above and on-ground features is similar, hence from this experiment we cannot conclude that the way the height data was used during segmentation has any positive effect. Interestingly the completeness/correctness of car-pixels is better by 2 to 3%, compared to SVL_1, especially the confusion with impervious surfaces is less in this experiment. This might be a hint that the height when included in the segmentation helps to reduce the merge of cars with the background.

Classification after a simple 5x5 chessboard segmentation (SVL_5, SVL_6)

As discussed in the segmentation section we also tested the classification after a simple chessboard segmentation. The results cannot be directly seen as a per-pixel classification, but some trends are probably the same. Concerning completeness and correctness of classification we see that the overall accuracy is similar to the other cases, but especially low vegetation and cars are much worse here compared to SVL_1: in the overall statistics the F1-score for cars is around 46% (here) vs. 57% (SVL_1). Many small objects, i.e. especially cars, are missed, else largely we can make the same observations: shadow areas are misclassified and the transition from above-ground to ground objects is often not well preserved.

Concerning delineation of objects we can see -- as expected -- the “rasterization effect” through the 5x5 pixel clusters. Especially in man-made objects with good separation to the background the delineation through image-based segmentation is of better quality. Also the CRF result (SVL_6) shows much less “speckle” effect than the boosting-only result from SVL_5, hence the observation from former experiments are confirmed. Refer to Figs 10 and 11 below for the results of SVL_5 and SVL_6, respectively. Interestingly, in the shown tile 8, the completeness for most classes is much better here than in SVL_1, but in the overall result, taking into account all validation tiles, the performance is comparable (expect for the cars and low vegetation, see above).

Discussion and Conclusion

The results show that the normalized height feature significantly contributes to the quality of classification, while the computation of vegetation indices is not really necessary. The CRF-extension helps to smooth the result, although the pixel-based evaluation might not be better as in the boosting-only result. To include the height for the segmentation seems to help only marginally in this case. However, it must be noted that the critical areas – transition from ground to off-ground objects are not always accurately represented in the employed normalized DSM. It would be interesting to perform the same experiments using the LIDAR DSM instead the matching DSM.

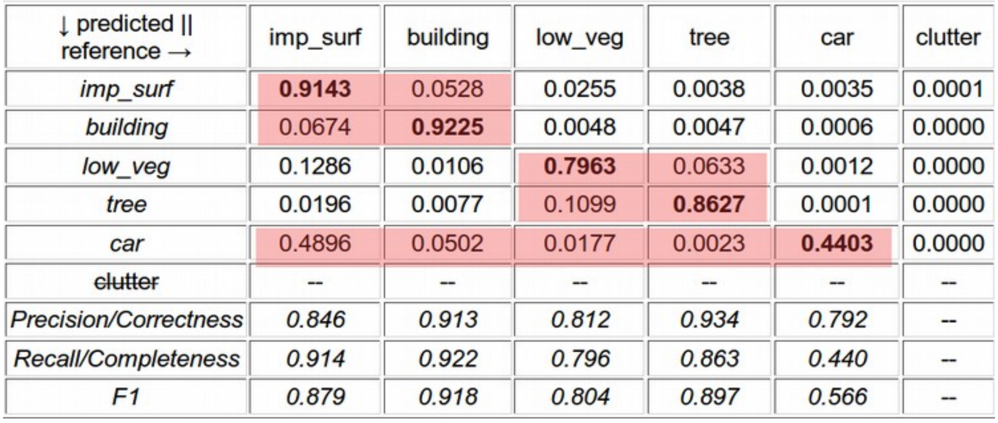
The fact that the problem cases (shadow, height transitions) are observable likewise in the image-based pre-segmentation and the chessboard segmentation shows that the largest impact comes from the features for classification, rather than the segmentation. Given that more small objects get missed in the chessboard classification case and that the computation times are much higher (by factor 12 in our case) an image-based pre-segmentation is advised for practical applications.

Some criteria concerning the relevance of a classification result have been defined by Mayer et al. (2006); they claimed a minimal completeness of 70% and a correctness of at least 85%. According to these the performed classification can be considered relevant for practical applications, at least for the classes impervious surfaces, trees and buildings. The correctness of low vegetation is 78%, hence below this threshold. Besides the already mentioned influence of the height on the classification result, one main problem is shadow casted on vegetated areas.

Availabilty of data and software

The normalized height images are available through Research Gate, http://www.researchgate.net/profile/Markus_Gerke, see the datasets attached to this technical report. **Attention!** The filtering was performed using the lasground tool in a batch process. There are still partly quite some large errors, e.g. in some tiles (like in tile 31) large industrial halls have been labeled as ground and are thus not included in the normalized DSM. Keep this in mind when using this dataset.

Pedro Felzenszwalbs segmentation code got modified in order to export the segments in txt format, as input for the SVL. It is available from the author of this report on request.



↓ predicted || reference →	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.9143**	0.0528	0.0255	0.0038	0.0035	0.0001
building	0.0674	**0.9225**	0.0048	0.0047	0.0006	0.0000
low_veg	0.1286	0.0106	**0.7963**	0.0633	0.0012	0.0000
tree	0.0196	0.0077	0.1099	**0.8627**	0.0001	0.0000
car	0.4896	0.0502	0.0177	0.0023	**0.4403**	0.0000
~~clutter~~	--	--	--	--	--	--
Precision/Correctness	*0.846*	*0.913*	*0.812*	*0.934*	*0.792*	--
Recall/Completeness	*0.914*	*0.922*	*0.796*	*0.863*	*0.440*	--
F1	*0.879*	*0.918*	*0.804*	*0.897*	*0.566*	--

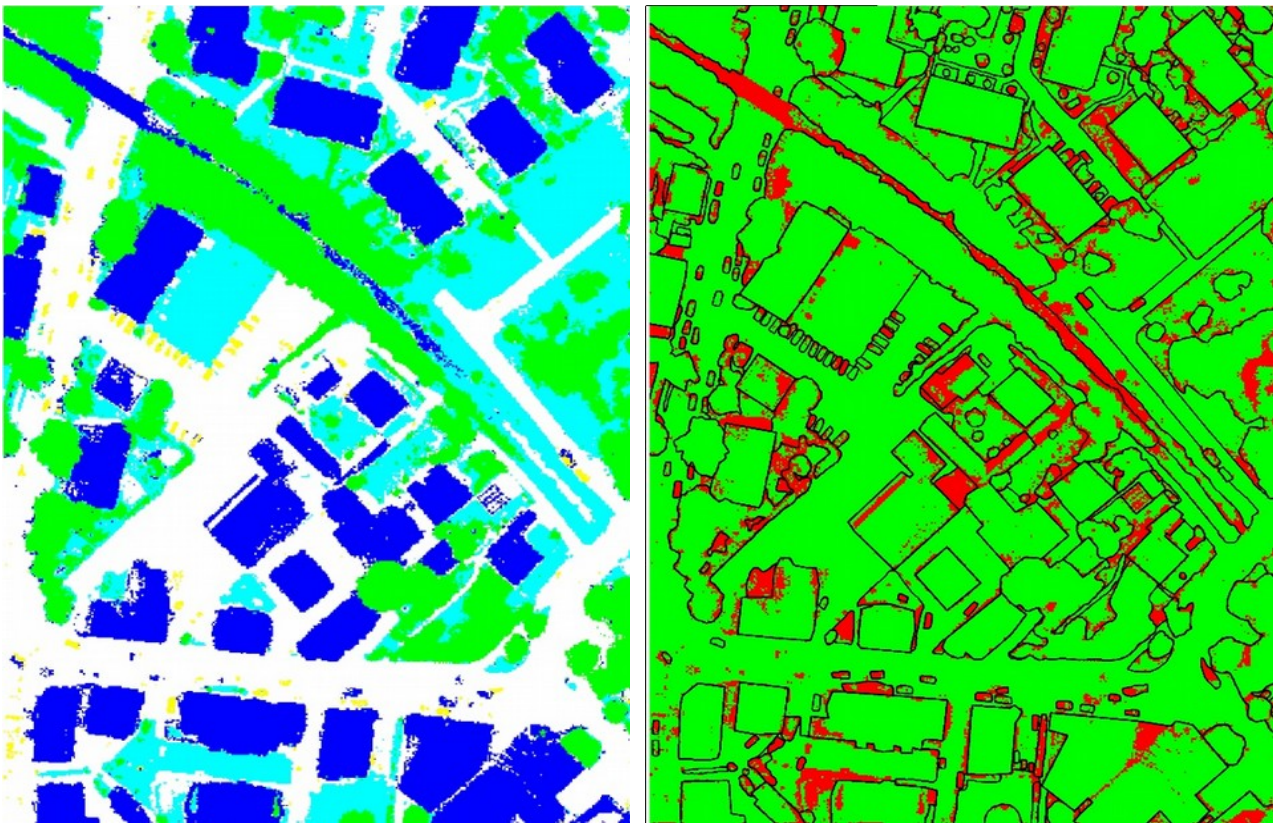
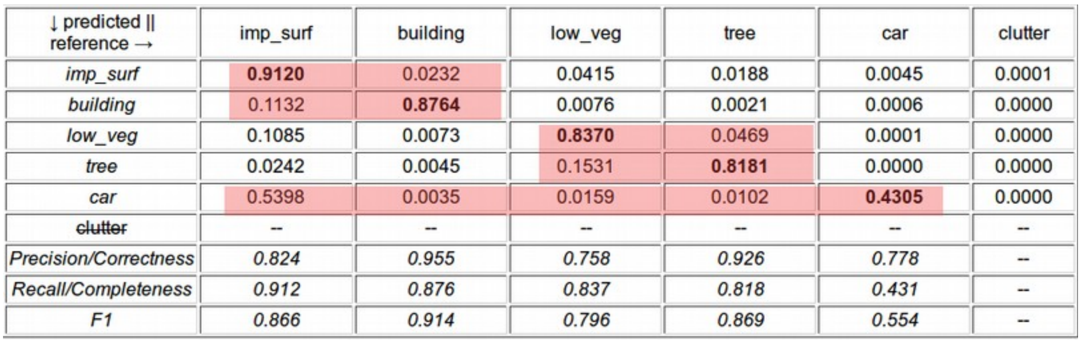


Figure 10: Results of SVL_5: same features and classification method as in SVL_3 (full features, no CRF), but segments from chessboard pre-classification



↓ predicted || reference →	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.9120**	0.0232	0.0415	0.0188	0.0045	0.0001
building	0.1132	**0.8764**	0.0076	0.0021	0.0006	0.0000
low_veg	0.1085	0.0073	**0.8370**	0.0469	0.0001	0.0000
tree	0.0242	0.0045	0.1531	**0.8181**	0.0000	0.0000
car	0.5398	0.0035	0.0159	0.0102	**0.4305**	0.0000
~~clutter~~	--	--	--	--	--	--
Precision/Correctness	*0.824*	*0.955*	*0.758*	*0.926*	*0.778*	--
Recall/Completeness	*0.912*	*0.876*	*0.837*	*0.818*	*0.431*	--
F1	*0.866*	*0.914*	*0.796*	*0.869*	*0.554*	--

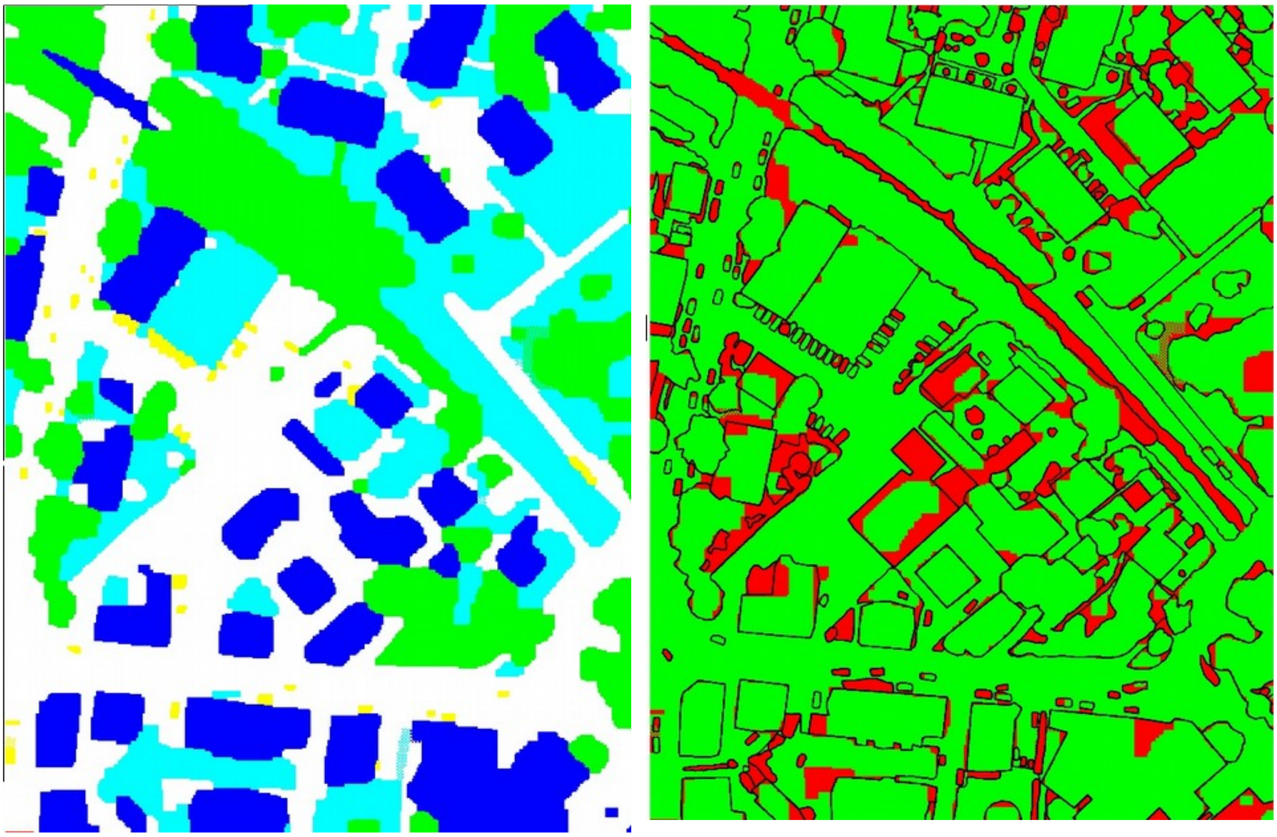


Figure 11: Results of SVL_6: same features and classification method as in SVL_1 (full features, with CRF), but segments from chessboard pre-classification

References

Axelsson, P., 2000. DEM generation from laser scanner data using adaptiveTIN models. In: ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. 23-4, pp. 110-117.

Felzenszwalb, P., and D. Huttenlocher, 2004. Efficient graph-based image segmentation, International Journal of Computer Vision, 59(2):167–181.

Gould, S., 2014. Stair Vision Library, URL: http://robotics.stanford.edu/~sgould/svl/ (last date accessed: 20 December 2014).

Gould, S., J. Rodgers, D. Cohen, G. Elidan, and D. Koller, 2008. Multi-class segmentation with relative location prior, International Journal of Computer Vision, 80(3):300–316.

Mayer, H., S. Hinz, U. Bacher, and E. Baltsavias, 2006. A test of automatic road extraction approaches. In: International Archives of Photogrammetry, RemoteSensing and Spatial Information Systems. Vol. 36-3, pp. 209–214.

Rottensteiner, F., G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction, ISPRS Journal of Photogrammetry and Remote Sensing, 93: 256–271.