Semantics through Time: Semi-supervised Segmentation of Aerial Videos with Iterative Label Propagation

Alina Marcu[1,2], Vlad Licaret[1], Dragos Costea[1,2], and Marius Leordeanu[1,2]

[1] University Politehnica of Bucharest, 313 Splaiul Independentei, 060042 Bucharest
[2] Institute of Mathematics of the Romanian Academy, 21 Calea Grivitei Street, 010702 Bucharest, Romania
{alina.marcu,vlad.licaret,dragos.costea,marius.leordeanu}@upb.ro

More details and qualitative results are shown below to further demonstrate the effectiveness of our proposed method. We show the impact of homography on SegProp, qualitative results before and after training SegProp, the impact of adding only the iterative algorithm on top of other methods and timing details. Additional video content is included alongside this document.

1 Discussion about convergence

Here we present the numerical performance of SegProp in plot form (Figure 1). The theoretical properties of our algorithm suggest convergence towards a singular value if enough iterations are computed, regardless of the starting point. What matters most is the static graph represented by optical flow and the ground truth data that is always forwarded unchanged on each iteration (see Section 2.1 in the paper). This progressive improvement can also be observed qualitatively in Figure 2.

2 Adding Homography to SegProp

Our method can support an arbitrary number of voting schemes. In the paper we present a homography based voting solution which we introduce after qualitatively assessing our initial results (Section 4.3). While Ruralscapes does not provide instance segmentations, we make the assumption that continuous labels are likely to correlate across a small enough interpolation distance Δt. Similarly, we assume that a mapping between two correlated regions can be approximated by a planar transformation for a sufficiently small Δt. We therefore identify connected components for each class map in P_i and P_j and project each such component onto P_k by estimating a homography transformation between flow based correspondences – we detail this method in Algorithm 1. In practice, we use a least-median robust method ($LMEDS$) for estimating H as a straight least squares derivation often fails for small objects due to the large number of outliers.

Our intuition is that such a mapping will help the labeling of moving objects and will better preserve the segmentation edges. We support this idea

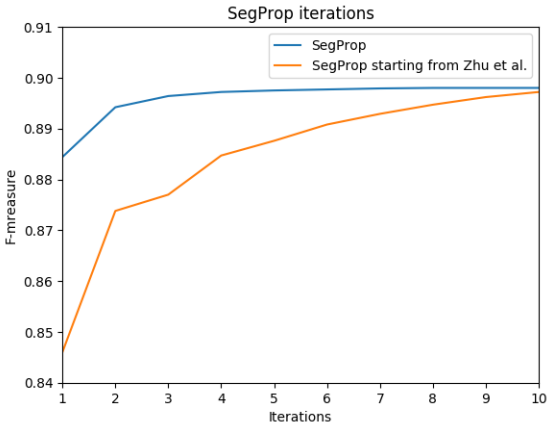

Fig. 1. The iterative aspect of our algorithm manages to improve segmentations even more. Even though [1] starts with poorer original segmentations, the iterations yield a better gain. However, having a good original segmentation helps as convergence should be achieved in fewer iterations.

with experimental results presented in the main paper (Table 3) and additional qualitative results shown in Figure 3. Replacing our flow-based votes with the homography mappings instead of using them together was also tested, but the numerical results suffer as not all connected component transformations can be satisfyingly estimated.

3 SegProp - Discussion

Majority voting. The final step of our algorithm is a simple majority vote - the class with the greatest cumulative score wins. However, it can happen that two or more classes share an equal maximum score - we estimate that approximately 0.12% of total pixels suffer from this class uncertainty at decision time on the first pass of SegProp, and this number naturally decreases as more votes are counted in future iterations. In our current implementation there is no special handling of this state, the first class is selected by the $max()$ function from an arbitrarily ordered array. Future work could include better handling of this edge case, for example by counting neighbouring votes or considering a class priority list.

Comparison with Zhu et al.[1]. While SegProp performs better both numerically and qualitatively for our use case, the method of Zhu et al. has at least one advantage over ours - the ability to propagate a single labeled frame, while SegProp requires a minimum of two. However, we achieve better results over larger time steps and on regions far away from the camera, at the cost of using an extra segmentation. Indeed, Zhu et al. [1] only use their method for rel-

Fig. 2. Qualitative results show the differences between competing methods, such as [1] and our algorithms. The iterations contribute to the performance improvement, resulting in better details and better edge alignment, especially for smaller objects.

Algorithm 1 Homography voting

1) Generate an additional voting map by computing homography transformations between connected components CC (connected regions with the same class label) from P_i and their flow based correspondences on frame k:

for each CC in P_i **do**
  for (x, y) in CC **do**
    $l_{i\to k}(x, y) = f_{i\to k}(x, y) + (x, y)$
  end for
  $H_{i\to k} \leftarrow LMEDS(CC, L_{i\to k})$
  for (x, y) in CC **do**
    $p^H_{i\to k}(x, y) = H_{i\to k}(p_i(x, y))$
  end for
end for

2) Repeat the first step for P_j and construct $P^H_{j\to k}$

3) Accumulate these two new votes with the first 4

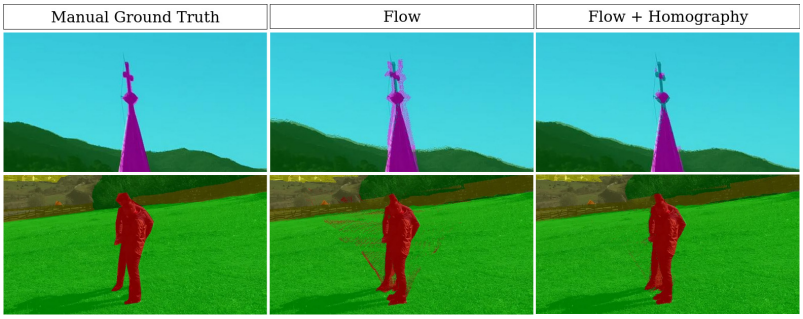

Fig. 3. The influence of adding the homography based mapping to our voting pool. Camera movement causes objects on different planes of reference to move one against the other. Both optical flow errors and imperfect segmentations make label propagation difficult in this case, but a structure-preserving homography proves useful. We show image crops focusing on details.

atively short distances of 1 to 5 frames, but for training purposes, segmentations might prove most useful when they are spaced further apart. Another advantage of their method is the increased computational efficiency compared to our full iterative approach. However, we still achieve both better results (see Figure 4) and faster running times using just one iteration (see Table 1).

About the complexity of our task and approach. Our aerial scenarios are in fact more difficult than many street-level car datasets. Ours has significant 6D pose changes (varied altitudes, viewpoints, rotations, 50kmph speed), varied and complex scenes, strong perspective effects and many different types of occlusions. The frame rate (50 fps) is high, but the number of propagated frames is also large. The actual propagation time is what matters most. Our algorithm is not simple in the way it uses iterative optical flow and homography voting, followed by 3D filtering. It is a form of spectral clustering, which is novel in video semantic segmentation literature. It is guaranteed to converge to the principal eigenvector of the space-time video graph, which ensures stability and global optimization under L2-norm constraints. That is the key reason why our SegProp, with different starting points, converges towards the same result.

4 Qualitative SegProp results, after training

Figure 5 presents more qualitative results for several state-of-the-art neural networks before and after training with our proposed method, SegProp. The first three rows show favourable results of our method compared to the baseline. The last three rows show the impact of the CNN choice in terms of performace - while the vanilla U-net and SafeUAVNet are similar, the former yields poorer results. DeepLabv3+ tends to fragment the labels, resulting in overall poorer segmentation.

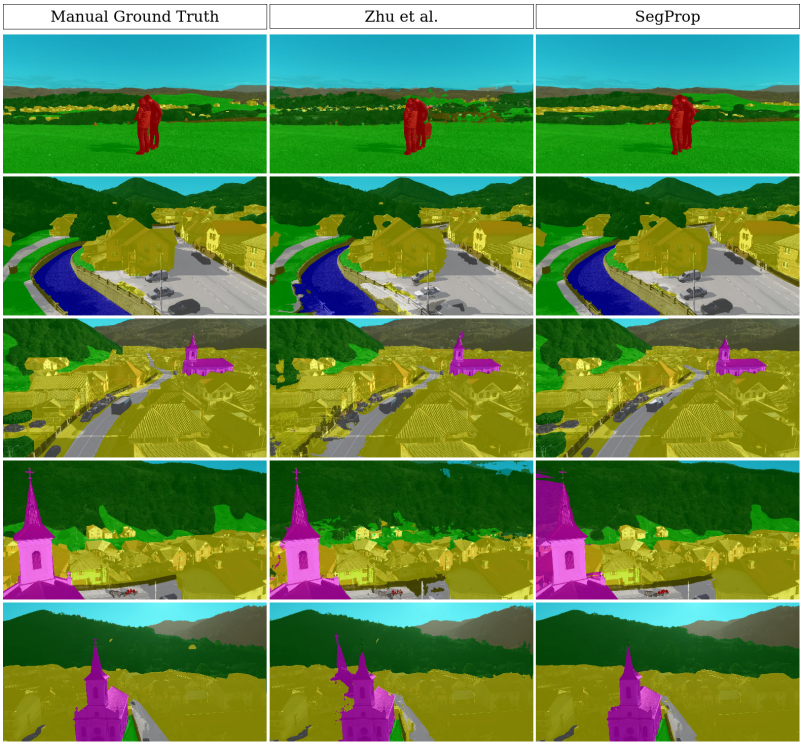

Fig. 4. SegProp compared to [1]. Better small object segmentation, better edges and more consistent detections are shown by SegProp. Heavily relying on optical flow and without a feedback mechanism, [1] tends to result in inconsistent labels - see the fence, land, water and church areas from the images above.

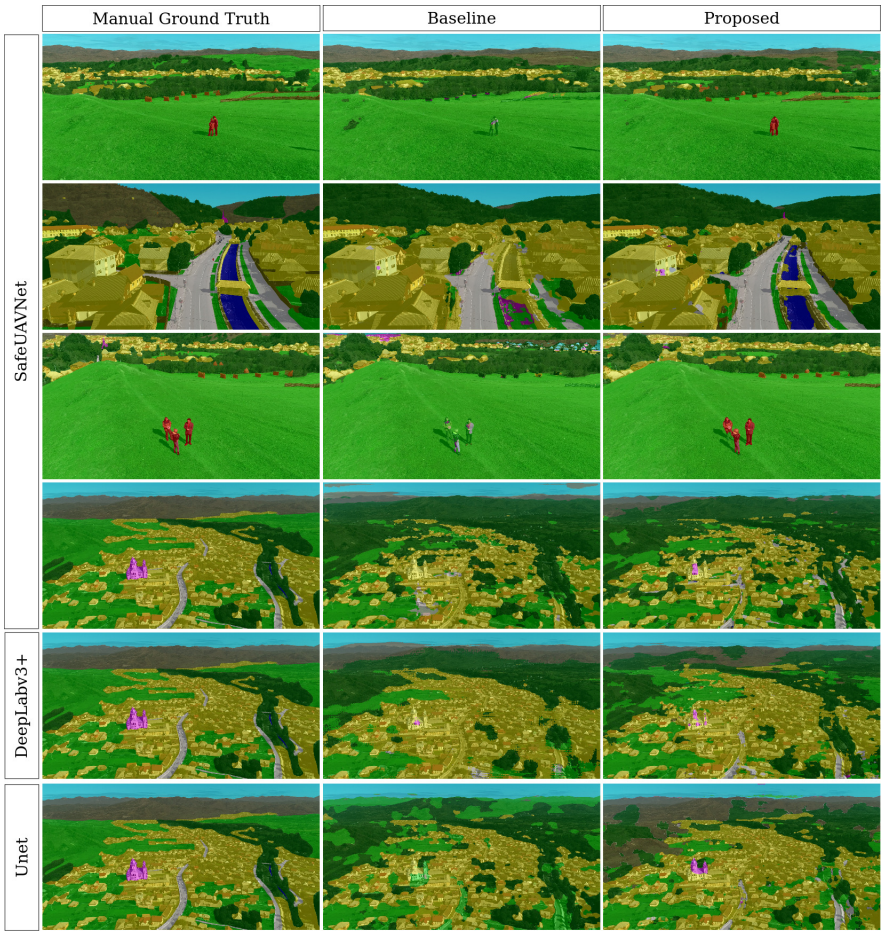

Fig. 5. Additional qualitative results on the testing set. Our method improves both the small and large object detection. For example, the humans in the third row are detected better, but also the skyline is more accurate (4th, 5th, 6th row). Even in uncertain label scenarios, such as the hill from the last row, our method yields a more plausible segmentation.

5 Timings

Table 1 presents the timing requirements of our method. While a single iteration is faster than [1], generating almost one frame per second at 2048×1080px, adding homography or iterations increases the computational cost. Nevertheless, there is a linear cost associated with the iterations - the algorithm can be stopped when timing constraints are reached.

Table 1. Timing results for one frame. The numbers below are computed for the rescaled images (2048×1080px).

Method	Runtime (seconds)
Zhu et al. [1]	1.74
SegProp Iteration 1	1.12
SegProp Iteration 1 + Homography	12
SegProp Iteration N, with $N > 1$	5.14

6 Videos

The video **TRAIN_segprop_vs_prediction.mp4** shows the main differences between our label propagation method before and after training the CNN. Although the quality of the interpolations produced over large time spans by our algorithm is reasonable, SegProp suffers from oscillations (jumps) when the ground truth is changed. Furthermore, the quality of the segmentation is strongly influenced by the quality of the optical flow. This might raise several problems, especially for small moving objects (such as persons or cars). In these cases, SegProp misses small objects completely. After training, the CNN learns from both the automatically labeled frames and also from manual ground truth (only 1%, which is a very small percent compared to the volume of the dataset) jointly, and is capable of improving the quality of the segmentations on top of SegProp. Even though the pipeline predicts each frame individually, it has smoother transitions, in addition to improving the detection of small objects.

The video **TEST_unseen_videos.mp4** shows qualitative results of the CNNs after training with SegProp, on unseen videos from the testing set. These results prove the generalization capabilities of our algorithm.

References

1. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 8856–8865