HD Maps: Fine-grained Road Segmentation by Parsing Ground and Aerial Images

Gellért Mártyus  
Remote Sensing Technology Institute  
German Aerospace Center  
gellert.mattyus@dlr.de

Shenlong Wang, Sanja Fidler, Raquel Urtasun  
Department of Computer Science  
University of Toronto  
{slwang, fidler, urtasun}@cs.toronto.edu

Abstract

In this paper we present an approach to enhance existing maps with fine grained segmentation categories such as parking spots and sidewalk, as well as the number and location of road lanes. Towards this goal, we propose an efficient approach that is able to estimate these fine grained categories by doing joint inference over both, monocular aerial imagery, as well as ground images taken from a stereo camera pair mounted on top of a car. Important to this is reasoning about the alignment between the two types of imagery, as even when the measurements are taken with sophisticated GPS+IMU systems, this alignment is not sufficiently accurate. We demonstrate the effectiveness of our approach on a new dataset which enhances KITTI [8] with aerial images taken with a camera mounted on an airplane and flying around the city of Karlsruhe, Germany.

1. Introduction

We are in an exciting time for computer vision, and more broadly AI, as the development of fully autonomous systems such as self-driving cars seems possible in the near future. These systems have to robustly estimate the scene in 3D, its semantics as well as be able to self-localize at all times. Key to the success of these tasks is the use of maps containing detailed information such as road location, number of lanes, speed limit, traffic signs, parking spots, traffic rules at intersections, etc.

Current maps, however, have been created with the use of semi-automatic systems that employ many man-hours of laborious and tedious labeling. An alternative to this costly labeling is to employ existing maps and correct/enhance them based on ground imagery or LIDAR point clouds, captured, for example, by a Velodyne/cameras mounted on top of a car. Systems like TESLA auto-pilot [1] are currently using their deployed fleet of cars, which are equipped with cameras, to perform such corrections. However, it is difficult to create full coverage of the world as we will need access to imagery/LIDAR from millions of cars in order to reliably enhance maps at a world-scale.

Alternatively, aerial images provide us with full coverage of a significant portion of the world, but at a much lower resolution than ground images. This makes semantic segmentation from aerial images a very difficult task. In this paper, we propose to use both aerial and ground images to jointly infer fine grained segmentation of roads. Towards this goal, we take advantage of the OpenStreetMap (OSM) project, which provides us with freely available maps of the road topology in the form of piece-wise linear road segments. We formulate the problem as energy minimization, inferring the number and location of the lanes for each road segment, parking spots, sidewalks and background, as well as the alignment between the ground and aerial images. We employ deep learning to estimate semantics from both aerial and ground images, and define a set of potentials exploiting these semantic cues, as well as road constraints, relationships between parallel roads, and the smoothness of both the estimations along the road as well as the alignment between consecutive ground frames.

We demonstrate the effectiveness of our approach in a new dataset which covers a wide area of the city of Karlsruhe in Germany, both from the ground and from the air. We provide pixel-level annotations for the aerial images in terms of fine-grained road categories. We call our dataset Air-Ground-KITTI. We show that our approach is able to estimate these categories reliably, while significantly reducing the alignment error between the ground and aerial images when compared to a sophisticated GPS+IMU system.

2. Related work

For several decades, researchers from various communities (e.g., vision, remote sensing) have been working on automatic extraction of semantic information from aerial images. In the following, we summarize the approaches most relevant to our work.
Aerial image parsing: Early approaches employed probabilistic models that aimed to produce topologically connected roads. [2] defined a probabilistic model that tiled the image into patches, performed road inference inside each patch via dynamic programming, and then “stitched” together high-confidence patches to ensure road connectivity. Recent work exploits learned classifiers to perform semantic segmentation. [15, 16] trained a neural net to classify pixels in local patches as road. They employ a post-processing step to ensure a consistent road topology across the patches, which is, however, prone to block-effects. [26] segments the road by defining an MRF on superpixels. High-order cliques are sampled over straight segments or junctions to encourage a road-like network structure. Due to complexity of high order terms a sampling scheme is used to concentrate on more important cliques. [4] samples graph junction-points using image consistency and shape priors. A full review of this large field is out of scope of this paper, and we refer the reader to [14] for a detailed review.

Aerial parsing with maps: While proven useful in many computer vision and robotics applications [9, 13, 3, 25], few works employ map information for parsing aerial images. [20] uses a screenshot of the vector map as a weak source of ground-truth for training a road classifier. [27] exploit road center-lines from OSM maps as a ground-truth road location and performs road segmentation by estimating the width of the road. This is done by finding boundaries of superpixels along the direction of the road, and ignoring dependencies across different line (road) segments. However, the alignment between OSM and aerial images is far from perfect. To solve this problem, [12] proposed a MRF which reasons about re-positioning the road centerline and estimating the width of the road. Smoothness is incorporated between consecutive line segments by encouraging their widths to be similar. In our work we go beyond this approach by introducing a formulation that reasons about more fine-grained road semantics such as lanes, sidewalks and parking spots, and exploits simultaneously aerial images as well as ground imagery to infer this information.

Fine-grained road parsing: Very few works exist that extract detailed segmentation. [17] propose a hierarchical probabilistic grammar to parse smaller-scale aerial regions into roads, buildings, vehicles and parking lots. Classifiers are first employed to generate object/building/vegetation proposals while the grammar imposes semantic and geometric constraints in order to derive the final parse. Learning and inference are both hard in grammars, and computationally expensive sampling techniques typically need to be employed. In our work, we are aiming at a detailed parsing of the roads into sub-categories. Unlike [17], we exploit OSM information in order to derive an efficient formulation. The work most related to ours is [21] which exploits the map as a screenshot of the road vector map to perform road and lane estimation. The authors take a pipeline approach, where, in the first step, road lane hypotheses are generated based on the output of the road classifier and detected lane markings. In the second step, the authors provide heuristics to “track” the lane hypotheses and connect them into a single lane labeling.

Aerial-to-ground reasoning: Recent work aims to exploit both aerial and the ground-view, mainly for the problem of geo-localization. In [11], a deep neural network is used to match ground images with aerial images in oblique views. The matches come from facade to facade matching and therefore can not be extended to orthographic aerial images. In [22], 3D reconstructions from the ground images are matched to oblique views of aerial images. [10] learn cross-view matching between ground images, aerial orthographic photos and land cover attributes. This extends the image geolocalization to areas not covered by ground images. Forster et al. [7] match the computed 3D maps of
MAVs and ground robots for localization and map augmentation. This method relies on matching 3D information and therefore needs multiview images both from above and on ground. In our work, we exploit the maps as well as ground and aerial imagery to perform fine-grained road parsing. We are not aware of prior work that tackles this problem.

3. Fine-grained Semantic Parsing of Roads

We now describe our model that infers fine-grained semantic categories of roads from aerial and ground images. In particular, we are interested in estimating sidewalks, parking, road lanes as well as background (e.g., vegetation, buildings). Towards this goal we exploit freely available cartographic maps (we use OSM), that provide us with the topology of the road network in the area of interest. Our approach takes as input an aerial image $x_A$, a road map $x_M$ and a set of ground stereo images $x_G$, which are taken by a calibrated stereo pair mounted on top of a car. The map $x_M$ is composed of a set of roads, where each road is defined as a piece-wise linear curve representing its centerline.

3.1. Model Formulation

We formulate the problem as the one of inference in a Markov random field (MRF), which exploits deep features encoding appearance in both aerial and ground images, edge information, smoothness in the direction of the road as well as restrictions between parallel roads to avoid double counting the evidence. Our model encodes each street segment in the aerial image with 15 random variables encoding all possible combinations of background (B), sidewalk (S), road lanes (L) and parking (P). In particular,

$$\mathbf{y} = (y_1, \ldots, y_{15})$$

$$= (B_1, S_1, B_2, S_2, P_1, L, P_2, S_3, B_3, S_4, B_4)$$

with $B_1, B_4$ the rightmost (leftmost) border of the background. We model roads with up to 6 lanes, i.e., $L = (L_1, L_2, L_3, L_4, L_5, L_6)$. We allow all variables (but $L_6$) to take the state of the previous random variable in the sequence (i.e., $y_i = y_{i-1}$), encoding the fact that some of these regions might be absent, e.g., there is no parking or sidewalk. This is not the case for $L_6$ forcing the fact that at least one lane should be present. We define the states of each random variable to be $[-15, 15]$m from the projection of the OSM centerline in the aerial image (Fig. 1). This discretization represents pixel increments. Note that while there are 15 random variables, $y$ defines 16 different regions as $B_1$ and $B_4$ are not limited on the left (right).

Each region width is simply defined by $w_i = y_i - y_{i-1}$, while the width of $B_1$ is defined as $w_1 = -15m + y_1$, and the width of $B_4$ as $w_{16} = 15m - y_{16}$, since $-15m$ and $15m$ are the beginning and end of the state space. Note that the combination $(B, S, B, S)$ is necessary (both on the left and right), as there are many bike lanes in Germany (where our imagery is captured), and it is not possible to distinguish them from the sidewalk. Fig. 1 illustrates the model.

Each of our ground images comes with a rough alignment with the aerial image as we have access to a GPS+IMU and the cameras are registered w.r.t these sensors. This alignment is, however, noisy with 1.67m error on average. Thus, our model reasons about the alignment when scoring the ground images. Towards this goal, we define $t = (t_1, \ldots, t_n)$ to be a set of random variables (one per ground image) representing the displacement in the direction perpendicular to the OSM road segment. We define the state space of each misalignment to be $t_i \in (−4m, 4m)$. This is discretized to represent pixel increments.

We define the energy of the MRF as to encode the information contained in the ground and aerial images as well as smoothness terms and constraints on the possible solutions:

$$E(y, t, x_A, x_M, x_G) = E_{\text{air}}(y, x_A) + E_{\text{ground}}(y, t, x_G) + E_{\text{smooth}}(y, t, x_M) + E_{\text{const}}(y)$$

(1)

We now define the potentials we employ in more detail.

Aerial semantics: We take advantage of deep learning in order to estimate semantic information from aerial images. In particular, we create pixel-wise estimates of 5 semantic categories: road, sidewalk, background, building and parking. We exploit the CNN for segmentation [23, 19] trained

Figure 2. BCD: The graph shows a simplified network with two parallel roads (each with 3 random variables) and one ground image per segment connected to the right road. BCD alternates between three types of updates. (a) Along the road updates: we optimize over each chain with the same color (while holding all other variables fix). The pairwise terms fold to unaries (see dashed black lines). (b) Perpendicular to the road updates: we do inference for the nodes with the same color (holding the rest fix). (c) Along the ground alignments: We minimize only the $t$ variables which are depicted in green. The $y$ variables are fixed and are depicted in black.
on ILSVRC-2014, which we fine-tune for a 5-label classification task: road, parking spot, sidewalk, building and background. To train the network we extracted training examples by extracting patches centered on the projection of the OSM road segments. If the road segment is too long (i.e., long straight road) we create an example every 20m. We further perform data augmentation by applying small rotations, shifts and flips to the training examples. The output of the softmax is a downsampled segmentation. To create our features, we upsample the softmax output using linear interpolation as in [5]. To save computation, we only apply the network in the region of interest (regions of the image that are close to OSM roads). The aerial semantic potential then encodes the fact that our final segmentation should agree with the semantics estimated by the deep net. Towards this goal, we define 5 features for each of our 16 regions, one per label of the deep net. Each feature simply aggregates the output of the softmax in that region. Recall that each region is defined by two consecutive random variables, e.g. the first sidewalk is defined by \( y_1, y_2 \), that is \( \sum_{i} x_i \), where \( x_i \) is the state at position \( i \). We refer the reader to Fig. 1 for an illustration. While this potential seems pairwise in nature, we can further decompose it into unary potentials via accumulators \( A \) perpendicular to the road direction. These are simply generalizations of integral images from axis aligned accumulators to accumulators over arbitrary directions. We thus define

\[
\phi_{i\ell}(y^i_1, y^i_{j+1}) = \sum_{p \in \Omega_i^\ell(y^i_1, y^i_{j+1})} \varphi(p) = A(y^i_{j+1}) - A(y^i_j)
\]

with \( y^i_j \) the \( i \)-th variable of the \( j \)-th road segment, and \( \varphi(p) \) the softmax output interpolated at pixel \( p \). To compute this features, we only need 5 accumulators per road segment, one for each semantic class that the deep net predicts.

**Aerial edges:** This potential encodes the fact that the location of the boundaries between regions should be close to image edges. We thus apply the edge detector of [6] to detect edges in our aerial images. We then define the potential to be the sum of the edges on the boundary between consecutive regions. To make it more robust we thicken the boundary to be of size 3 pixels.

**Along the road smoothness:** We encode smoothness along the road by encouraging consecutive road segments to be similar. In particular, we use the \( \ell_1 \) distance between consecutive road estimations in the direction of the road, i.e.

\[
\phi_{\text{sm}}(y^i_1, y^i_{j+1}) = |y^i_j - y^i_{j+1}|
\]

**Parallel roads:** The regions of close by parallel roads can overlap. To avoid double counting the evidence, we incorporate an additional constraint that forces \( S_1 \) of the second road to be bigger or equal to \( B_4 \) of the first road or vice versa. We refer the reader to Fig. 1 for an illustration.

**Road collapse constraints:** We force each variable \( y^i_j \) to have a state higher or equal than the previous variable, so that the order is preserved. Note that equal means that a road can collapse (i.e., does not exist)

\[
\phi_{\text{coll}}(y^i_j, y^i_{j+1}) = \begin{cases} 
\infty & \text{if } y^i_{j+1} < y^i_j \\
0 & \text{otherwise}
\end{cases}
\]

The only exception is \( L_6 \), which we force to have non-zero width as otherwise we could have a road segment without road. Thus

\[
\phi_{\text{ex}}(L_5, L_6) = \begin{cases} 
\infty & \text{if } L_6 \leq L_5 \\
0 & \text{otherwise}
\end{cases}
\]

**Lane size constraint:** This constraint forces each region, if present, (i.e., if it is not taking state 0) to have a minimal and maximal size. In particular, we use (1m-3m) for sidewalk, (1.8m-4.5m) for parking and (2.3m-4.6m) for each road lane. Note that width 0 is allowed so that regions can disappear if they are not present in the road segment (e.g.,
we only have two lanes, there is no sidewalk on the highway). The intervals for the lanes are estimated based on the standards of German roads, while the sidewalk and parking intervals are computed based on empirical estimates.

Centerline prior: As our images are well registered with OSM, we include a prior that the centerline of our model should be close to the centerline of OSM. In particular,

$$\phi_{cen}(L_3) = \begin{cases} \|L_3 - l\|_2 & \text{if } -7.5 \leq L_3 \leq 7.5 \\ \infty & \text{otherwise} \end{cases}$$

with $l$ the location of the centerline.

Ground semantics: We take advantage of deep learning in order to estimate semantic information from ground images. We exploit the VGG [23] implementation of [19] trained on PASCAL VOC, which we fine-tuned to predict the same 5 classes as the aerial semantics (road, parking, sidewalk, building and background). We estimate the ground plane from the stereo image and project pixels belonging to this plane to the aerial image via a homography. We then define our ground semantic potential to encourage the segmentation to agree with the aligned ground image segmentation projected to the aerial image. Towards this goal, we define 5 features for each of our road regions, each counting the amount of softmax output for the given class:

$$\phi_{ground}(t_k, y^1_i, y^{1}_{i-1}) = \mathcal{G}(t_k, y^1_i) - \mathcal{G}(t_k, y^1_i)$$

Note that via the integral accumulator the 3-way potential decomposes into pairwise terms $\mathcal{G}(t, y)$. In this case we
null
Algorithm 1 Block coordinate descent inference (BCD).

1: Set all alignments \( t = 0 \), and initialize \( y \) by minimizing Eq. (1) ignoring the along road smoothness.
2: repeat
3:  for all \( y^j \) do
4:      Minimize Eq. (1) along the road w.r.t \( y^j \), holding the rest fixed.
5:  end for
6:  for all \( y_i \) at one segment of the road do
7:      Minimize Eq. (1) w.r.t \( y_i \), holding the rest fixed.
8:  end for
9:  for all \( t \) variables do
10:     Minimize Eq. (1) w.r.t \( t \), holding \( y \) fixed.
11: end for
12: until no energy reduction or max number iterations

\( \ell_{sm} = |y_{i}^{j} - y_{i}^{j+1}| \). Note that the geometrical constraints in our model are either 0 or \( \infty \) and are not trained.

4. Experiments

We collected a new dataset which we call Air-Ground-KITTI, which is composed of both ground images from the KITTI tracking benchmark [8] and newly acquired orthorectified aerial images over the same area. We neglected the KITTI sequences where the car is mostly static, resulting in 20 KITTI sequences for a total of 7603 ground stereo images. We annotated every 30th ground image with 4 semantic classes (parking, sidewalk, road, building). The aerial images were acquired by a DSLR camera mounted on an airplane and projected on the earth surface with 9 cm/pixel Ground Sampling Distance (GSD). We split the data into 10 training and 10 test aerial image/KITTI sequences, with special care to avoid overlaps in the aerial images. We manually annotated the aerial images with 4 categories (parking, sidewalk, road, building) as closed polygons and the lane markings as polylines. This took 70h of annotation, at a mean of \( 21h/km^2 \), the area is \( 3.23 \, km^2 \).

To perform fine-grained segmentation using both aerial and ground images, we estimate a homography that transforms the ground plane in KITTI to the UTM coordinate system based on the KITTI’s GPS+IMU measurements and the camera calibration. We assign each ground image to the closest parallel road segment. Our model then refines this estimate in the direction perpendicular to each road segment. We process every 5th ground image in the sequence.

As metrics for the fine-grained segmentation we calculate the pixelwise Intersection over Union (IoU), Precision, Recall and F1 metrics for three classes (i.e., road, parking, sidewalk). Note that we only measure the areas laying in the area of interest (i.e., \( \pm 15m \) around the road map centerline). We consider two parallel roads overlapping over the same area as a serious error. To reflect this, we handle these areas as if they were background. The metrics in Table 1 are calculated according to this.

For the roads, we additionally compute whether we have estimated the correct number of lanes. This is measured as the average \( \ell_1 \) error in terms of number of lanes (EN). Note that if there are no lane markings, estimating the number of lanes is very difficult. Fig. 7 (a-b) shows this difficulty.

Figure 7. It is hard to estimate the number of lanes if there are no lane markings. (a) Our method, (b) Oracle (i.e., our method with ground truth potentials). (c) The OnlyLane model without the parallel constraint allows the road to "jump" to the nearby parallel road. (d) The parallel constraints of LaneRoadParallel prevents this from happening. (e) Dense, urban pedestrian streets are difficult to estimate. (f) Our model is not intended for intersections, as it does not reason about turn lanes.

In our experiments, we compare our approach to the state-of-the-art method of [12], which uses OSMs to estimate road width. We also tested different model configurations for our approach. We refer to Lane as a model that employs Aerial semantics, Aerial Edges, Road collapse constraints, Lane size constraint and Centerline prior energy terms. Inference is done independently for each road segment via dynamic programming along the \( y^t = y^1_t, \ldots, y^{15}_t \) chains. We refer by LaneParallel to a model where we additionally include the constraint between nearby parallel road. We refer by LaneRoad as a model that contains all the potentials in Lane plus smoothness along the road. We apply BCD inference by alternating between the chains perpendicular to the road (the lanes) and along the roads (segments). We refer by LaneRoadParallel a model that contains all potentials but the ground. Finally, Ground contains all potentials. We evaluate this case only where ground images are available.
Table 1. Performance for the semantic classes (i.e. road, parking spot, sidewalk) with various models and the two baselines. The values are in %, except EN which is the average road lane number $l_1$ error with respect to the oracle. * Marks the method of [12] with our deep road classifier. The last two rows marked with ** evaluate only over areas where ground images are also available.

Table 2. Ground to air image misalignment based on the camera calibrations (GPS+IMU) and after our alignment measured in meters. Using ±4 meter interval.

Comparison to the state-of-the-art: As shown in Table 1, our method outperforms [12] in almost all metrics, even when we apply our deep features instead of their road classifier in their method. Furthermore, we retrieve more semantic categories such as sidewalk, individual road lanes and parking. The constraint between parallel roads is important to achieve good results on roads. Without it, our model cannot outperform [12], which has this constraint.

Deep semantic features in aerial Images: We show the performance of our Deep Network in Fig. 4. Note that it is much better than the road classifier of [12].

Alignment between aerial and ground images: As shown in Table 2 and Fig. 3 reasoning about the alignment between ground and aerial images while doing fine-grained segmentation improves the alignment significantly.

Qualitative Results: We visualize our results when using only aerial images in Fig. 5, and when using joint aerial and ground reasoning in Fig. 6. Our approach is able to estimate well the lanes, sidewalk and parking as well as the alignment between the ground and the aerial images.

Ablation studies: As shown in Table 1, the metrics for different versions of our model are fairly similar, however qualitatively, as we add more potentials, the results get better. This is illustrated in Fig. 7 (c), where the OnlyLane model moves the middle road to a parallel road resulting in a noncontinuous structure. In contrast, the LaneRoadParallel model prevents overlaps and favors smoothness, see the Fig. 7 (d). Including the ground images only slightly improves performance. We believe this could be overcome by using stronger features in the ground images, i.e., leveraging the full 3D point cloud, not just the ground plane. Note that since our approach gives us very precise alignments between the ground and the aerial images it could be used to enhance OSM with object locations, e.g. traffic signs.

Inference time: Inference in our full model takes 6 seconds per km of road, with a single thread on a laptop computer. Note that BCD can easily be parallelized.

Limitations: Our model is designed for individual roads and it does not reason about turning lanes connecting different roads at intersections (see Fig. 7 (f)). Dealing with such scenarios is part of our future work. Semantic segmentation from aerial images reasons mainly about the visible parts of the street. Therefore covered areas (e.g. by building, bridges, trees) can be a problem. However, when ground images are available, our approach can handle this problem. Our aerial images were acquired in early spring, and thus trees occluding the roads is not a big problem.

5. Conclusion

We proposed an approach to enhance existing freely available maps with fine grained segmentation categories such as parking spots and sidewalk, as well as the number and location of road lanes. Towards this goal, we proposed an efficient method that produces very accurate estimates by performing joint inference over both, monocular aerial imagery captured by a plane and ground images taken from a stereo pair mounted on top of a car. We have demonstrated the effectiveness of our approach on a new dataset which enhances KITTI with aerial images taken with a camera mounted on an airplane and flying around the city of Karlsruhe. In the future, we plan to reason about other fine grained categories such as traffic signs in order to further enhance the maps. As our method reasons about the accurate alignment between the map and the ground images, we envision its use for precise, lane-wise self localization of the vehicle on the road.

Acknowledgments

We would like to thank Viktoria Zekoll, Stefan Turzer and Márk Bagdy for making the laborious annotation work.
References

[27] J. Yuan and A. Cheriyadat. Road segmentation in aerial images by exploiting road vector data. In COM.geo, 2013. 2