Global alignment of HD maps and satellite imagery for Autonomous Navigation

Srirangan Madhavan, Harish Rithish and Manoj Kilaru

UCSD
smadhavan@ucsd.edu
hsethuram@ucsd.edu
mkilaru@ucsd.edu

1 Introduction

1.1 Motivation

Autonomous driving is one of the most important thrust areas for application of AI and computer vision. Datasets in this space are created by leveraging sets of highly detailed maps that are developed partly with manual annotation. These are meticulously annotated and provide information about various lanes, signs and traffic rules, speed limits, etc. Such datasets exist for a few cities and provide accurate annotations. However, this method consumes a lot of resources to create, update while also not being tenable to develop at scale. While these datasets are useful, they do not provide a Birds Eye View of the road which has been shown to be highly valuable for navigation. Aerial images spatially cover the entire world and consume far less resources for acquisition. We could use these aerial images along with the HD maps provided by autonomous driving companies [3] to train the deep learning models for automating the tasks such as semantic segmentation. However, aerial images are acquired from a different source and thus, have to be accurately aligned with the existing HD maps before they can be used for autonomous navigation. Alignment issues between these two sources are mainly due to HD maps being created by assuming a flat-view of the world while satellite imagery is taken from a spherical Earth. In this work, we seek to minimize the alignment error between satellite imagery and HD maps by formulating it as an energy minimization problem [18].

1.2 Related Works

There are two parts to the pipeline that we use here. The first one being image segmentation and the the other being the main problem of image alignment itself. These problems have been tackled with various approaches in the past few decades. Those approaches can be broadly summarized as follows.

Earlier approaches of image segmentation were based on thresholding. These methods selected a globally optimal threshold by maximizing the inter-class variance. Later region growth based segmentation were used. The methods start with seed pixels and iteratively grow the region of pixels similar to them in

the neighbourhood. Edge detection based segmentation methods and clustering based segmentation have also been successfully used in various cases. Latest advancements came with the advancement in neural networks in the form of CNNs. Various methods by Badrinarayanan et. al. [1],Ronnenberger et. al. [14], Chen et. al. [4] have been developed that produce highly accurate segmentation results.

Specifically in the case of aerial image segmentation, the initial approaches to this problem were using probabilistic methods that had topological constraint for producing connected road segments [2]. In recent years many deep learning based methods have been developed [12][11][17]. A highly related past work on this is by Seo et. al.[15]. They use a road classifier and lane detection modules followed by a heuristic approach to connect the hypotheses generated into a single lane segment. Some other methods also apply leaning based pipelines to find correspondences between ground images taken by on board sensors and map them to a projected version of aerial maps and vice versa [8][16][6].

1.3 Contribution and Organization

In this assignment, we have adapted pretrained model from Zhou et al. [19] to get segmented roads as outputs from aerial images. The major part of this assignment is to rectify the alignment differences between nuscenes dataset and the aerial images obtained from google maps so that the aerial maps can then be further used for autonomous driving applications. The approach we are taking here is formulating this as an optimization problem and using first a simple search algorithm and then a descent method called Block Coordinate Descent (BCD) to reach the minimum. The novelty and main attraction for this method is that it is practically unsupervised and does not require the resources required by learning based methods in terms of compute power or data. Further sections of this report are divided into a section describing the primal and dual of optimization problem formulation, the dataset and workflow and the approach taken to solve the problem.

2 Problem Statement

We take as input, HD maps and their corresponding satellite imagery. Since these two pieces of data are obtained from different sources, there is significant misalignment due to which they cannot be used together for downstream processing. In this work, we present an efficient method to obtain alignment through Block Coordinate Descent optimization.

We achieve this by formulating the problem as a convex optimization problem. Let us define the cost of alignment as the required shift(displacement) in GPS co-ordinates. Displacement is a 2-d vector one dimension corresponds to shift in latitude and other longitude. Let us define the penalty for any misalignment in the form of an energy function $E(I_1, I_2, d)$ where I_1 and I_2 are the aerial image and HD map images and d is displacement. For this problem, we

constraint on the variable displacement. we want the absolute value of displacement in both longitude and latitude to be less a scalar. We will go through the energy function which we minimize in detail.

Energy Function

Unary Potential: For each displacement we first compute overlap for each block between HD map and displaced aerial image as number of pixel matches. We want to maximize this overlap, but we want to minimize the energy, so we define unary potential as weighted sum of overlap and displacement as follows

$$\phi(I_1, I_2, d) = \sum_i \sum_j W_1(1 - O(I_{1,ij}, C(I_{2,ij}, d_{ij}))) + W_2(d_{ij})$$

Where C is shift function and O is overlap function.

$$O(I, I') = \frac{\sum_x \sum_y I_{xy} * I'_{xy}}{\sum_{xy} I_{xy}} \quad \text{and} \quad C(I, d)_{xy} = I_{x-dx, y-dy}$$

Smoothness Constraint: We try to minimize difference between displacements of adjacent blocks.

$$S(I_1, I_2, d) = \sum_i \sum_j \|d_{i+1,j} - d_{i,j}\| + \|d_{i,j+1} - d_{i,j}\|$$

So combining them gives us energy function in BCD.

$$E(I_1, I_2, d) = W_3 * \phi(I_1, I_2, d) + W_4 * S(I_1, I_2, d)$$

Optimization

We need to compute optimum d that minimizes energy function.

$$d^* = arg\, min_d E(I_1, I_2, d)$$

Primal

We need to compute optimum d that minimizes energy function.

$$\begin{aligned} min_d \quad & E(I_1, I_2, d) \\ \text{s.t.} \quad & \forall_{ij} |d_{ij}| \preceq c \end{aligned} \tag{1}$$

Dual

First we will compute dual without smoothness constraint. As there is no global constraint over all blocks we could just optimize each block individually. So I will drop ij in the notation and primal becomes

$$\begin{aligned} min_d \quad & W_1(1 - O(I_1, C(I_2, d))) + W_2(d) \\ \text{s.t.} \quad & d \preceq c \quad and - d \preceq c \end{aligned} \tag{2}$$

so lagrangian

$$L(d, \lambda_1, \lambda_2) = W_1(1-O(I_1, C(I_2, d))) + W_2(d) + \lambda_1(d-c) + \lambda_2(-d-c)\ \lambda_1, \lambda_2 \geq 0$$

so dual $g(\lambda_1, \lambda_2) = \inf_d L(d, \lambda_1, \lambda_2)$ differentiating W.R.T d_x

$$-W_1 \frac{\partial O(I_1, C(I_2, d))}{\partial d_x} + W_2 + \lambda_1 - \lambda_2 = 0$$

$$-W_1 \frac{\sum_i \sum_j I_{1,ij} * \frac{\partial C(I_2,d)_{ij}}{\partial d_x}}{\sum_{ij} I_{1,ij}} + W_2 + \lambda_1 - \lambda_2 = 0$$

Let I_{2X} be x direction derivative of image I_2 which can be computed using convolution with x derivative kernel. we know $\frac{\partial}{\partial d_t} f(x-t) = -1 * f'(x-t)$ from taylor series. so above equation reduces to

$$W_1 \frac{\sum_i \sum_j I_{1,ij} * I_{2X,i-d_x,j}}{\sum_{ij} I_{1,ij}} + W_2 + \lambda_1 - \lambda_2 = 0$$

So finally it can be written as.

$$W_1(O(I_1, C(I_{2X}, d))) + W_2 + \lambda_1 - \lambda_2 = 0$$

It is hard to write closed form solution this d_x let it be d^*.

So dual problem is

$$\begin{aligned} max_{\lambda_1,\lambda_2} \quad & W_1(1 - O(I_1, C(I_2, d^*))) + W_2(d^*) + \lambda_1(d^* - c) + \lambda_2(-d^* - c) \\ \text{s.t.} \quad & W_1(O(I_1, C(I_{2X}, d^*))) + W_2 + \lambda_1 - \lambda_2 = 0 \\ & d^* \preceq c \quad and - d^* \preceq c \quad and\ \lambda_1, \lambda_2 \geq 0 \end{aligned} \tag{3}$$

Similarly we can write for d_y This is the optimization we did in greedy search explained below and it doesn't have global constraints which we introduced by adding smoothness constraint and we solved global optimization using block gradient descent. The dual for global optimization will have a term for derivative of smoothness in f_0 and h_0 term of dual and there would double summations for some terms. we are not writing dual for primal to global as it is similar to above dual just following term would be the addition.

$$\begin{aligned} \frac{\partial}{\partial d_{ij,x}} S(I_1, I_2, d) = & -\frac{d_{(i+1,j),x} - d_{ij,x}}{\|d_{i+1,j} - d_{i,j}\|} - \frac{d_{(i,j+1),x} - d_{ij,x}}{\|d_{i,j+1} - d_{i,j}\|} + \frac{d_{ij,x} - d_{(i-1,j),x}}{\|d_{i,j} - d_{i-1,j}\|} \\ & + \frac{d_{ij,x} - d_{(i,j-1),x}}{\|d_{i,j} - d_{i,j-1}\|} \end{aligned}$$

KKT Conditions

let d^* and $(\lambda_1^*, \lambda_2^*)$ be any primal and dual optimal points with zero duality gap. So KKT conditions are as follows

$$\begin{aligned} \forall_{ij}\ d_{ij}^* - c\ &\leq 0 \\ \forall_{ij}\ -d_{ij}^* - c\ &\leq 0 \\ \lambda_1^*\ &\geq 0 \\ \lambda_2^*\ &\geq 0 \\ \forall_{ij}\ \lambda_1^*(d_{ij}^* - c)\ &= 0 \\ \forall_{ij}\ \lambda_2^*(-d_{ij}^* - c)\ &= 0 \\ \forall_{ij}\ \frac{\partial}{\partial d_{ij}} E(I_1, I_2, d^*) + \lambda_1^* - \lambda_2^*\ &= 0 \end{aligned} \tag{4}$$

Where

$$\frac{\partial}{\partial d_{ij,x}} E(I_1, I_2, d^*) = W_3 * \frac{\partial}{\partial d_{ij,x}} \phi(I_1, I_2, d^*) + W_4 * \frac{\partial}{\partial d_{ij,x}} S(I_1, I_2, d^*)$$

$$\frac{\partial}{\partial d_{ij,x}} \phi(I_1, I_2, d^*) = W_1(O(I_1, C(I_{2X}, d^*))) + W_2$$

3 Dataset and Workflow

We use the NuScenes [3] dataset for this project. The NuScenes dataset is widely used for the development of novel algorithms in the autonomous navigation space. The dataset contains a wide-suite of sensors, including LiDAR, RADAR and street-view images. In addition, the dataset contains a centimeter accurate human annotated semantic map, known as the NuScenes HD map. While satellite imagery have been shown to be useful for navigation, the NuScenes dataset, like many others in this space do not contain aerial imagery as part of its data suite. In this work, we present a generic method to augment autonomous navigation datasets with aerial imagery. We download aerial imagery from Google Maps at a zoom level of 21, fully encompassing the scope of this dataset. Since the NuScenes dataset and satellite imagery are obtained from different sources, they are not in alignment and cannot be directly used. We use the HD maps of NuScenes dataset to align the aerial imagery with the dataset. The NuScenes dataset contains four HD maps, 3 in Singapore and 1 in Boston. We apply the same alignment method for the sub-datasets and report our results. We note that each HD maps spans an area of 2Km×2Km.

Instead of directly aligning HD maps and aerial imagery, we first extract roads from aerial imagery using image segmentation. For this purpose, we then use the pre-trained model from Zhou et al. [19]. This model is trained on high-resolution (1024×1024) satellite imagery for road extraction. We use the roads extracted from aerial imagery as a proxy to align with the corresponding HD map. Figure

1 summarizes our workflow. We note that road segmentation methods are inaccurate and incomplete, making the optimization problem a challenging one to solve.

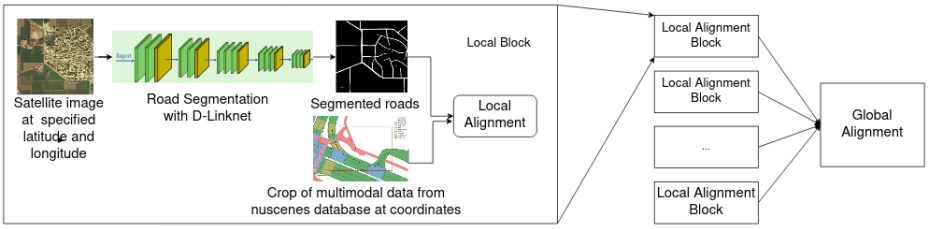


Fig. 1: The figure represents the workflow of our approach. We take aerial imagery and obtain the predicted roads through image segmentation. We obtain an initial local alignment between the predicted roads and HD maps through a Greedy approach. We then perform Block Coordinate Descent optimization for global alignment.

4 Approach

For each HD Map in NuScenes we have 64 Aerial images we want to align. We have two approaches for aligning which we go through in this section. Briefly in greedy approach we align 64 blocks independently and in second approach we do a smoothness globally by adding constraints on adjacent blocks.

4.1 Simple greedy search

In case of the greedy search, each of the block is allowed a search range of -15m to 15m in either directions(latitude and longitude), where this search range corresponds to the displacement of the block. For each block, the unary potential is given a weight factor of $1(W_1)$ and the displacement is given a weight factor of $0.001(W_2)$. Each block is aligned with its corresponding nuscenes map such that the bottom left corner is the south west corner of map. Then for each block, the overlap and the unary potential are calculated for each step in the search range and the configuration with the highest overlap (therefore least potential) and least displacement is chosen.

This method is faster and has much less computational requirement than the BCD algorithm. But this does not ensure continuity of road segments and is less accurate. Which we discuss in the results section.

Although this is less accurate, the output from greedy search can be used as a good starting point for the BCD algorithm to help with faster convergence.

4.2 Global Optimization using Block Coordinate Descent

We need to compute an optimum d that minimizes the energy function.

$$d^* = arg\,min_d E(I_1, I_2, d)$$

Inspired by [9], we perform an approximate minimization using Block Coordinate Descent(BCD) algorithm. As shown in Figure 2, each iteration of BCD comprises of four steps. Each step involves methodically restraining a group of blocks while optimizing the other blocks. Specifically, we fix even rows, odd rows, even columns and odd columns in the four steps. We initialize the displacement for each block with the results of greedy search. Post initialization, each step seeks to optimize the overall energy function, i.e., both the fixed and free blocks are included in the optimization function. So energy decreases after every iteration. We stop the algorithm when the difference in energy between any two adjacent iterations is below a user-specified threshold.

4.3 Computational Complexity

Computational complexity: $O(Ale/S)$, where A is the size of the map, l is length of the local search grid, e is the number of iterations and S is the size of an individual block.

4.4 Implementation Details

We provide the implementation details for only the Singapore-onenorth to keep the report succinct. The center of the HD map is [1.288210, 103.784751] in lat-long coordinates. One meter displacement in latitude and longitude is measured as $4.28 * 10^{-4}$ degrees. The local grid search space is $15m$ in each direction, with a step size of $1m$. The image resolution for each block of aerial imagery and HD map is 1024×1024. The maximum number of iterations is 5. The weights for overlap, local displacement, and smoothness are chosen empirically as 1, 0.001, and 0.03 respectively.

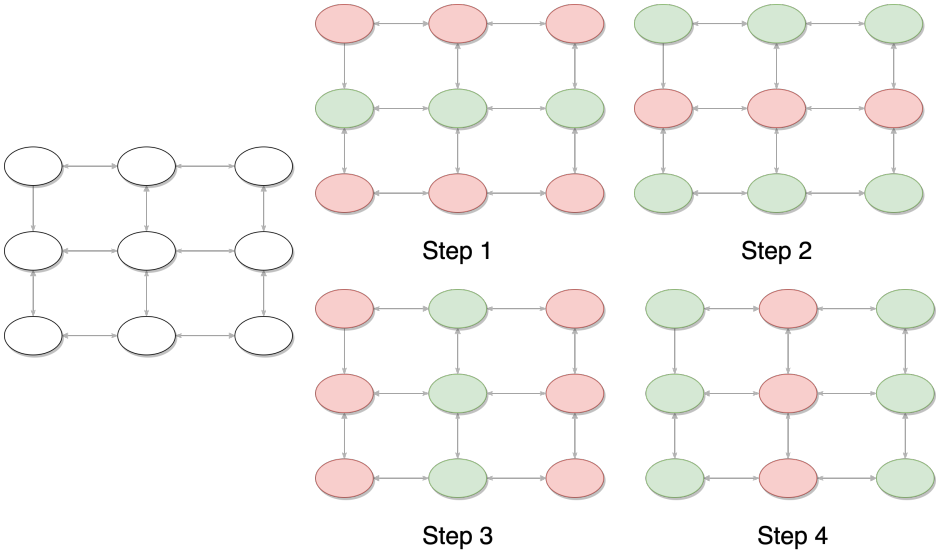


Fig. 2: This shows a single iteration of BCD. In each step we fix the red blocks and optimize only for green blocks

5 Results

We first validate our model quantitatively by measuring the overlap and smoothness, between the Greedy and BCD optimization. In Table 1, we present the percentage of overlap between the NuScenes HD map and the aerial imagery. We use the predicted roads on aerial imagery as a proxy to quantify the overlap with HD maps. We witness 30% increase in overlap after BCD optimization in comparison to the overlap before alignment. However, in comparison with the Greedy optimization, we notice a 2% decrease in overlap. This is expected as Greedy considers only local alignment and ignore global smoothness. This hypothesis is validated in Table 2, where we observe Greedy to have an undesirably high smoothness error of 7.94m. Through BCD optimization, we are able to considerably decrease the smoothness error to 2.56m.

We then observe display the outputs of our optimization method. Figure 3 compares the original NuScenes HD map with the Greedy and BCD stitched maps after alignment. The adjacent blocks are stitched first left to right and then top to bottom. In case of overlap, the latest placed block is shown on top. We notice that the stitched BCD map closely resembles the origingal NuScenes HD map. While not apparent, the Greedy approach contains many artefacts, while the BCD approach removes them through smoothing. To highlight the impact of smoothing, we present short segments of the HD map in 4. On the top row which represents the results of the Greedy methods, we notice the roads to be disconnected. On the bottom row, we notice that BCD optimization has managed to correct this artefact and make the roads continuous.

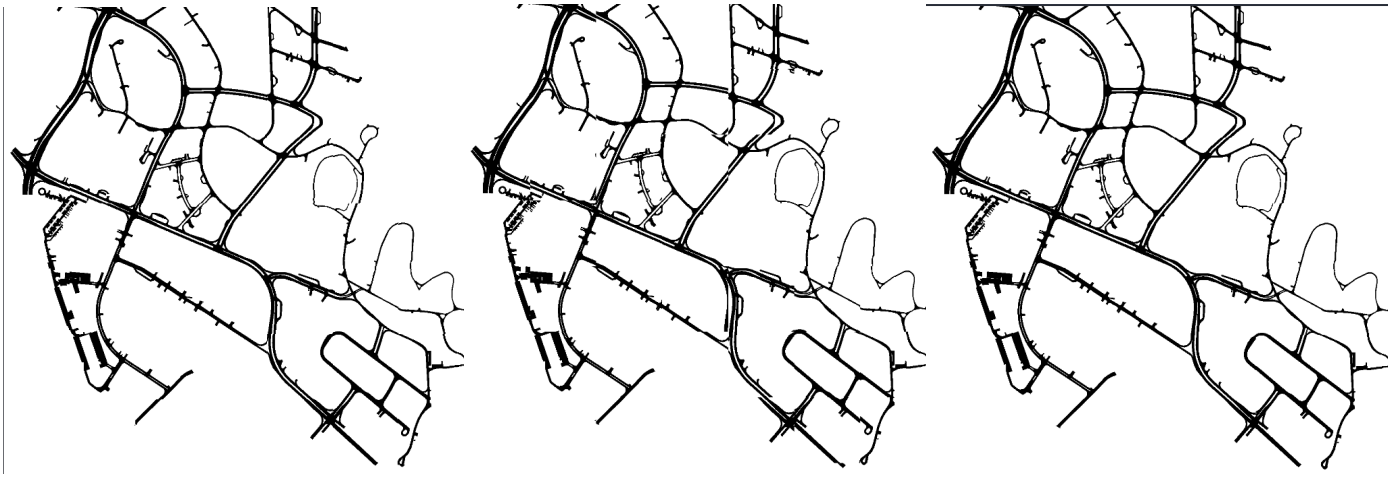
Additionally, we compare the overlap of roads in HD maps and aerial imagery through imposing the road layers one above the other, as shown in Figure 5. The first row displays the imposed map segments before alignment while the second row shows the map segments after alignment. It is clear that the overlap between HD maps and aerial imagery has significantly improved post BCD optimization. In Figure 6, we also display the imposition of the NuScenes road layer on top of the aerial imagery. We witness a significant alignment between these two layers. All these results show significant improvement than before alignment, validating the quality of our approach.

Dataset	Initial	Greedy	BCD
Boston	57.40%	85.63%	82.37%
Singapore 1	49.05%	79.41%	77.1%
Singapore 2	43.79%	74.29%	72.15%
Singapore 3	51.35%	81.27%	80.90%
Mean	50.39%	80.15%	78.13%

Table 1: % Overlap

Dataset	Greedy	BCD
Boston	6.53m	1.47m
Singapore 1	8.13m	2.93m
Singapore 2	9.27m	3.56m
Singapore 3	7.83m	2.31m
Mean	7.94m	2.56m

Table 2: Smoothness error



(a) NuScenes HD Map (b) Stitched image from greedy output (c) Stitched image from BCD output

Fig. 3: Comparison of the original NuScenes HD map with the Greedy and BCD stitched maps after alignment.

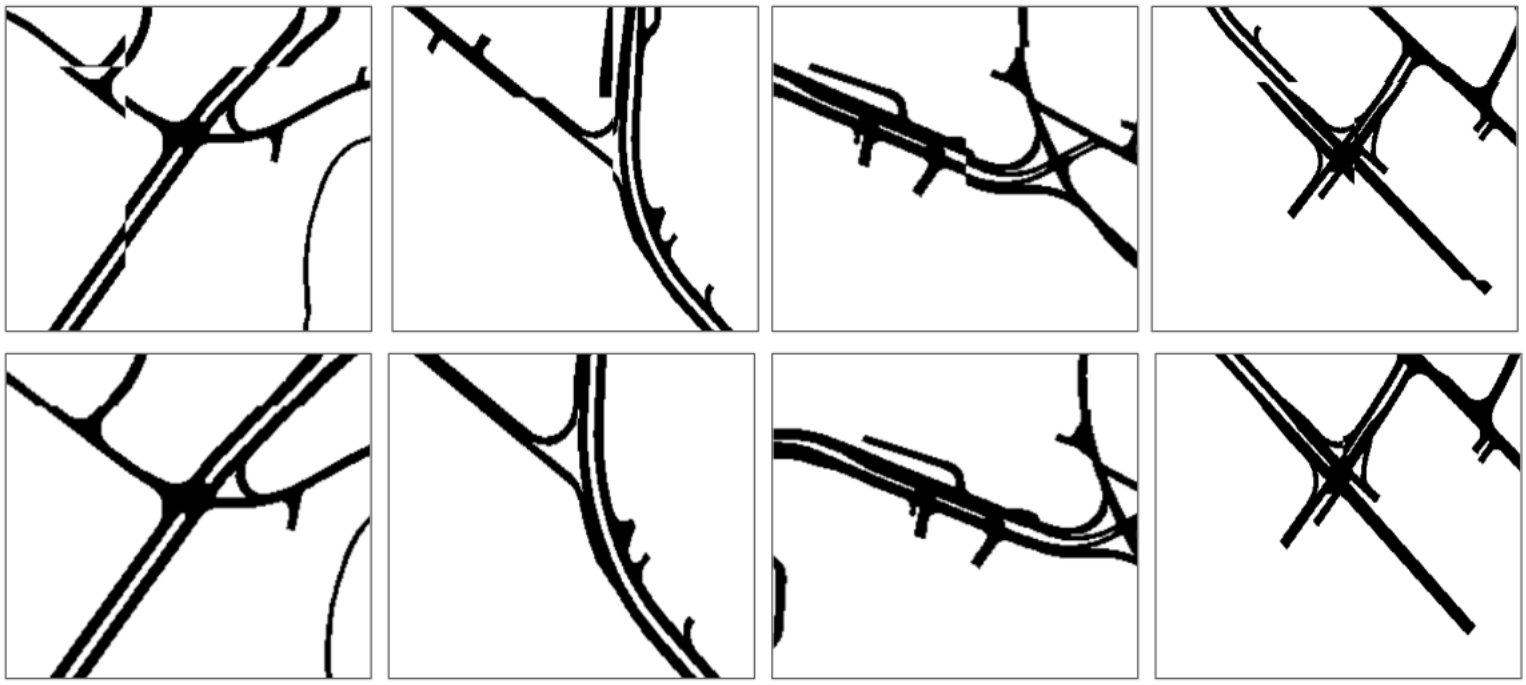


Fig. 4: Comparison of the cropped segments of the stitched HD map between Greedy (top row) and BCD (bottom row) approaches. We can clearly see artefacts in the Greedy approach while they are smoothed in the BCD outputs.

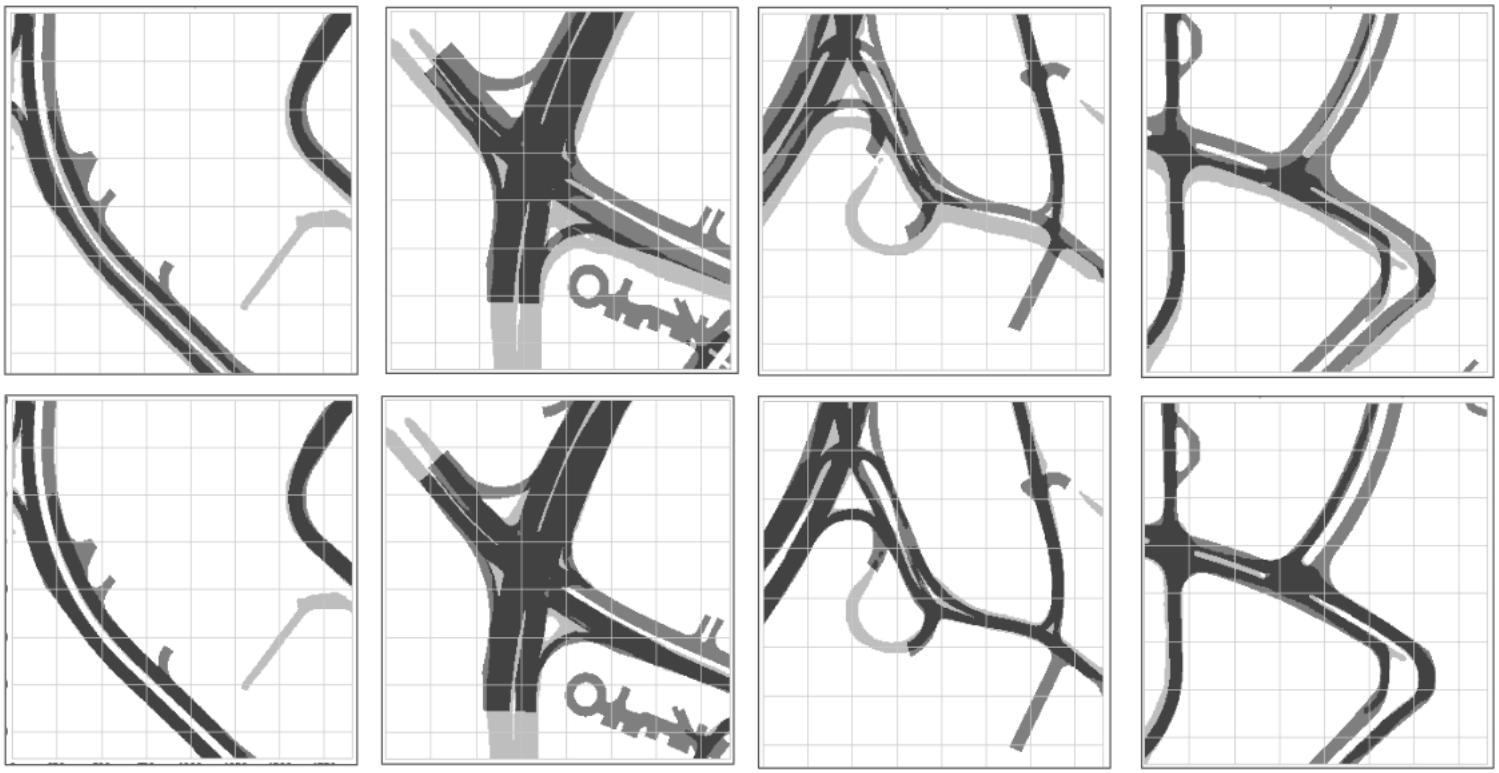
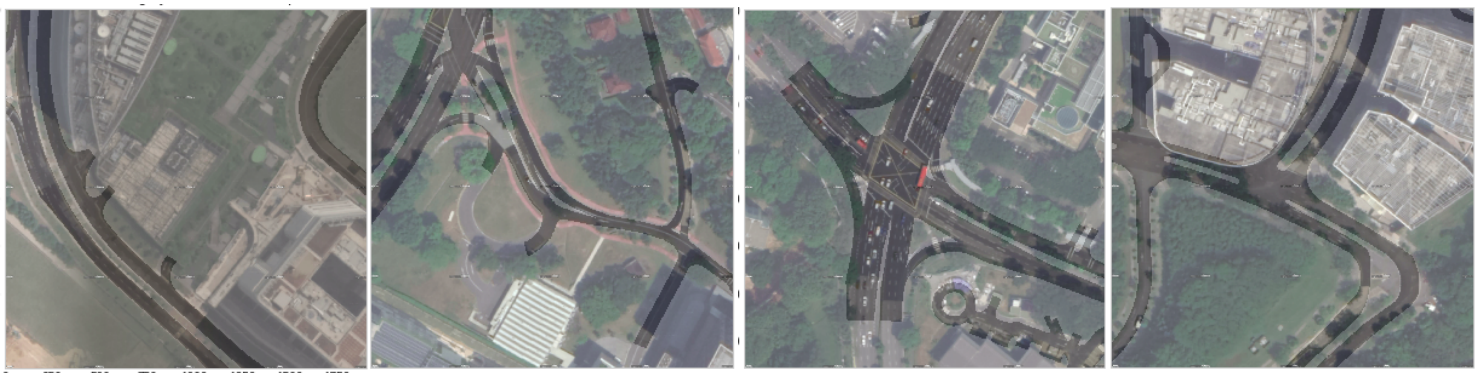


Fig. 5: In first row we have initial imposition of the roads in HD maps and aerial images. In the second row we show the imposition of the roads in HD maps and aerial images after alignment.



(a) (b) (c) (d)

Fig. 6: HD map and aerial imagery after alignment. We notice the roads in HD maps to clearly align with the roads in aerial imagery.

6 Conclusions and Further work

In this work, we are proposing an unsupervised, optimization based method that can be used to find the accurate correspondence between freely available aerial images and well annotated datasets like nuscenes. This line of work will enable the wide spread adaption of readily available data from satellite imagery in autonomous driving and reduce the cost of labour. The method that we have developed here is based on classical methods and is much less resource hungry compared to recent learning based approaches. In this work, we have mainly focused on utilizing the unary potential with a very relaxed constraint on the smoothness of the image. As part of improving this work, we would like to add more factors to the minimization problem that take into account other untapped features such as lane sizes, geometric constrains and semantics. Further we hope to move beyond iterative methods like block coordinate descent and find a more robust method that provides the speed of closed form solutions while utilizing the maximum of available data.

7 Task Assignment

The writing and preparation of report was split equally among the three members. Srirangan Madhavan conducted the initial literature survey before discussion. Manoj Kilaru derived and analyzed the dual and KKT of the greedy and block coordinate descent problems and implemented initial code for preparing data. Srirangan implemented the greedy solution and Harish Rithish implemented the main BCD algorithm and conducted the experiment with datasets and forming the results.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. CoRR **abs/1511.00561** (2015), http://arxiv.org/abs/1511.00561
2. Barzohar, M., Cooper, D.B.: Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. IEEE Trans. Pattern Anal. Mach. Intell. **18**, 707–721 (1996)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
4. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. CoRR **abs/1606.00915** (2016), http://arxiv.org/abs/1606.00915
5. Feng, T., Truong, Q.T., Nguyen, D.T., Koh, J.Y., Yu, L.F., Binder, A., Yeung, S.K.: Urban zoning using higher-order markov random fields on multi-view imagery data. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 614–630 (2018)

6. Forster, C., Pizzoli, M., Scaramuzza, D.: Air-ground localization and map augmentation using monocular dense reconstruction. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3971–3978 (2013). https://doi.org/10.1109/IROS.2013.6696924
7. Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K.: Learning aerial image segmentation from online maps. IEEE Transactions on Geoscience and Remote Sensing **55**(11), 6054–6068 (2017)
8. Lin, T.Y., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5007–5015 (2015). https://doi.org/10.1109/CVPR.2015.7299135
9. Mattyus, G., Wang, S., Fidler, S., Urtasun, R.: Enhancing road maps by parsing aerial images around the world. In: Proceedings of the IEEE international conference on computer vision. pp. 1689–1697 (2015)
10. Máttyus, G., Wang, S., Fidler, S., Urtasun, R.: Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3611–3619 (2016)
11. Mnih, V., Hinton, G.E.: Learning to detect roads in high-resolution aerial images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010. pp. 210–223. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
12. Mnih, V., Hinton, G.E.: Learning to label aerial images from noisy data. In: ICML (2012)
13. Máttyus, G., Wang, S., Fidler, S., Urtasun, R.: Enhancing road maps by parsing aerial images around the world. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1689–1697 (2015). https://doi.org/10.1109/ICCV.2015.197
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015), http://arxiv.org/abs/1505.04597
15. Seo, Y.W., Urmson, C., Wettergreen, D.: Ortho-image analysis for producing lane-level highway maps. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems. p. 506–509. SIGSPATIAL '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2424321.2424401, https://doi.org/10.1145/2424321.2424401
16. Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M.: Accurate geo-registration by ground-to-aerial image matching. In: 2014 2nd International Conference on 3D Vision. vol. 1, pp. 525–532 (2014). https://doi.org/10.1109/3DV.2014.69
17. Wegner, J.D., Montoya-Zegarra, J.A., Schindler, K.: A higher-order crf model for road network extraction. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1698–1705 (2013). https://doi.org/10.1109/CVPR.2013.222
18. Zampieri, A., Charpiat, G., Girard, N., Tarabalka, Y.: Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 657–673 (2018)
19. Zhou, L., Zhang, C., Wu, M.: D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 182–186 (2018)