Automatic 3D Railway Track Extraction and Reconstruction with Kalman Filter in Multiple Aerial Images

Dong Wei[a,1], Xiaotong Li[a,1], Yongjun Zhang[a,*], Chang Li[b] and Ziqian Huang[a,1]

[a]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430072, P.R.China

[c]College of Urban and Environmental Science, Central China Normal University, Wuhan, 430072, P.R.China

ARTICLE INFO

Keywords:
3D line segments
line clustering
line triangulation
3D line evaluation

ABSTRACT

Three-dimensional (3D) lines require further enhancement in both clustering and triangulation. Line clustering assigns multiple image lines to a single 3D line to eliminate redundant 3D lines. Currently, it depends on the fixed and empirical parameter. However, a loose parameter could lead to over-clustering, while a strict one may cause redundant 3D lines. Due to the absence of the ground truth, the assessment of line clustering remains unexplored. Additionally, 3D line triangulation, which determines the 3D line segment in object space, is prone to failure due to its sensitivity to positional and camera errors.

This paper aims to improve the clustering and triangulation of 3D lines and to offer a reliable evaluation method. (1) To achieve accurate clustering, we introduce a probability model, which uses the prior error of the structure from the motion, to determine adaptive thresholds;

1. Introduction

Currently, the length of the railway has exceeded 1.3 million kilometers on the earth, for which the maintenance and development of railways have a significant impact on safe operations. As the preliminary stage of extracting 3D railway track (RT) accurately and efficiently, to support engineering design, monitor construction quality, and ensure operational safety, has become one of the basic components in the maintenance of existing railways.

The extraction of RT can be achieved by real-time kinematics, LiDAR, and multiple images. The real-time kinematic is generally mounted on a railway measurement vehicle and obtains the RT by moving along the rail track. In general, it has a satisfactory accuracy while requiring operations on the track, thus demanding the cooperation of railway departments, and there are issues related to both safety and efficiency. LiDAR sensors can be mounted on a drone, which is more convenient and secure than real-time kinematic. Because a further process, like point segmentation or classification, is required for RT extraction, the drone must maintain a low flight altitude to satisfy the standards of the point-cloud density, which would impact the efficiency. A drone with cameras can capture aerial images efficiently with a safe distance from the railway area. But RT extraction is challenging in aerial images: (1) The dense points reconstructed with aerial images are inaccurate around the railway track because of the occlusion and matching problems caused by the parallax variation. (2) Joining image semantics to obtain RT might be workable; However, how to detect the semantics of RT accurately and completely in aerial images remains to be studied.

If we reconstruct the dense points from multiple aerial images and detect the RT from point clouds, the overall method of finding RT is similar to deal with the point clouds that obtained from mobile laser scanning (MLS) or airborne laser scanning (ALS). Generally, the RT can be detected with semantic segmentation, while the significant noise, inaccurate edge localization, and large density variations of point clouds bring about great challenges to the robust semantic segmentation. Thus, most general segmentation algorithm cannot be used directly in RT segmentation; instead, the carefully designed geometric priors was used to guide the segmentation and the grouping of RT: such as constructing the shape features and density data on the basis of railway bed extraction. However, these methods relies heavily on the quality and density of the point cloud, thus requiring the drone to maintain a low flight path to improve point cloud quality and reduce the processing range. Compared with point clouds, images contains rich

*Corresponding author
weidong@whu.edu.cn (D. Wei); zhangyj@whu.edu.cn (Y. Zhang); lichang@ccnu.edu.cn (C. Li)
[1]Co-first authors.

semantic informations. Thus, several studies exploited the deep learning method that design the network for training and detect the RT from aerial images, which demonstrated the effectiveness of deep learning technology in RT extraction. Moreover, the deep learning method relies heavily on training samples and considering the texture of railway regions varies greatly across the world, it may require an increased number of training samples to obtain a more generalizable detection network. In addition, these methods just deal with single frame and lacked the strategy of processing multiple aerial images.

This paper propose the accurate RT extraction for multiple aerial images, which fully exploits the contexture and geometry informations across multiple images:

- To exploit the geometry constraint of RT across multiple images, we extract the straight line from images as the basic geometry cell, for which we propose the robust clustering and triangulation methods. We first propose the noise-resistant clustering across multiple images to obtain the complete and non-redundant 3D line; Then, we propose the novel and accurate triangulation algorithm to refine the 3D line position of the RT.
- To exploit the rich texture information in images, our clustering method exploit the deep features of existing network trianed from millions of images, rather than a new network specifically designed for RT extraction; thus requiring non pre-training, which generaly needs expensive samples.

Compared to LiDAR based methods, we use more affordable imaging drones to conduct an efficient and safer railway aera maping than ALS drones or MLS equipments, and the rich contexture is exploited to compensate for issues caused by point cloud quality. Compared to the former image-based methods, we propsose the complete clustering and reconstruction strategies that obtain the accurate and non-redundant 3D RT from multiple aerial images; and non pre-training is required due to the utilization of geometry guidance in multiple images.

2. Related works

3. Methodology overview

The flow of our methodology is presented in Figure 1. We take aerial images of the railway area as input, for which SfM and point clouds are used in advanced with existing software. We first extract the image line and convert it to a space line with the locally optimal plane of the point cloud. Then, we cluster the single 3D line to RT candidates with the frame work derived from DBSCAN, during which the texture information of multiple images extracted from ResNet is used as one of the inlier distance. Having obtained the RT cluster, we then trace and reconstruct the vector-based RT in the Kalman framwork, which fully exploits the RT structure and the multi-view geometries to resolve the uncertainty caused by initial image line segment extraction and the point cloud error. The railway-track pair is the start seed of our Kalman method. We first convert image lines to 3D lines with the local optimal plane of the point cloud. Then, we cluster the single 3D line to RT pairs with DBSCAN frame work, during which the deep feature of multiple images is used as the inlier distance.

4. Kalman filter-Based railway-track reconstruction

In this study, we utilize the Kalman Filter to optimize the three-dimensional coordinates and directions of two points. The state vector is defined to encapsulate the positions and directions of both points, represented as a 12-dimensional vector:

$$\mathbf{x} = \left[\mathbf{p}_1, \mathbf{d}_1, \mathbf{p}_2, \mathbf{d}_2\right]^T \tag{1}$$

where $\mathbf{p}_i = \left(x_i, y_i, z_i\right)$ denotes the position and $\mathbf{p}_i = \left(dx_i, dy_i, dz_i\right)$ represents the direction. The state transition model assumes that at each time step t, each point moves in its respective direction by a fixed distance t, while maintaining a constant direction. The state transition matrix $\mathbf{F}$ is constructed as a block diagonal matrix comprising two identical 6×6 submatrices $\mathbf{F}_p$:

$$\mathbf{F}_p = \begin{bmatrix} \mathbf{I}_3 & t \cdot \mathbf{I}_3 \\ \mathbf{0}_3 & \mathbf{I}_3 \end{bmatrix}, \quad \mathbf{F} = \text{blkdiag}(\mathbf{F}_p, \mathbf{F}_p) \tag{2}$$

where $\mathbf{I}_3$ is the 3×3 identity matrix and $\mathbf{0}_3$ is the 3×3 zero matrix. The process noise covariance matrix $\mathbf{Q}$ accounts for uncertainties in position and direction and is similarly structured:

$$\mathbf{Q}_p = \text{diag}(q_p, q_p, q_p, q_d, q_d, q_d), \quad \mathbf{Q} = \text{blkdiag}(\mathbf{Q}_p, \mathbf{Q}_p) \tag{3}$$

Here, q_p and q_d represent the process noise variances for the position and direction components, respectively. The observation model integrates both actual measurements of the points' positions and directions and the imposed constraints. The observation matrix $\mathbf{H}$ is augmented to include these constraints:

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_{12} \\ \mathbf{H}_c \\ \mathbf{H}_{dr} \end{bmatrix} \tag{4}$$

where $\mathbf{I}_{12}$ is the 12×12 identity matrix corresponding to the direct measurements of the state vector. The constraint matrices are defined as:

$$\mathbf{H}_c = \begin{bmatrix} \mathbf{0}_{3\times3} & \mathbf{I}_3 & \mathbf{0}_{3\times3} & -\mathbf{I}_3 \end{bmatrix}, \quad \mathbf{H}_{dr} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0}_{3\times9} & -\mathbf{I}_3 \end{bmatrix} \tag{5}$$

The matrix $\mathbf{H}_c$ enforces direction consistency by computing the difference between the directions of the two points:

$$\mathbf{H}_c\mathbf{x} = \begin{bmatrix} dx_1 - dx_2 \\ dy_1 - dy_2 \\ dz_1 - dz_2 \end{bmatrix} \tag{6}$$

Similarly, $\mathbf{H}_{dr}$ imposes relative distance consistency by calculating the relative positions:

$$\mathbf{H}_{dr}\mathbf{x} = \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \\ z_1 - z_2 \end{bmatrix} \tag{7}$$

The observation noise covariance matrix $\mathbf{R}$ incorporates both the measurement noise and the noise associated with the constraints:

$$\mathbf{R} = \text{blkdiag}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_c, \mathbf{R}_{dr}) \tag{8}$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ correspond to the measurement noise of the first and second points, respectively:

$$\mathbf{R}_1 = \text{diag}(r_p, r_p, r_p, r_d, r_d, r_d), \quad \mathbf{R}_2 = \text{diag}(r_p, r_p, r_p, r_d, r_d, r_d) \tag{9}$$

$\mathbf{R}_c$ and $\mathbf{R}_{dr}$ represent the noise associated with the direction consistency and relative distance constraints:

$$\mathbf{R}_c = \sigma_c^2 \cdot \mathbf{I}_3, \quad \mathbf{R}_{dr} = \sigma_r^2 \cdot \mathbf{I}_3 \tag{10}$$

Here, σ_c and σ_r are the standard deviations controlling the strictness of the direction consistency and relative distance constraints, respectively.

The Kalman Filter operates through iterative prediction and update steps. In the prediction step, the state vector and covariance matrix are projected forward using the state transition model:

$$\mathbf{x}_{k+1|k} = \mathbf{F}\mathbf{x}_k \tag{11}$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}\mathbf{P}_k\mathbf{F}^T + \mathbf{Q} \tag{12}$$

In the update step, both the actual measurements and the constraints are incorporated. The combined observation vector $\mathbf{z}$ includes the actual measurements and the desired constraint outcomes (typically zero, indicating no difference for constraints):

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_{\text{actual}} \\ \mathbf{0}_{6\times1} \end{bmatrix} \tag{13}$$

The observation residual $\mathbf{y}$ is computed as:

$$\mathbf{y} = \mathbf{z} - \mathbf{H}\mathbf{x}_{k+1|k} \tag{14}$$

The residual covariance $\mathbf{S}$ and the Kalman Gain $\mathbf{K}$ are then determined:

$$\mathbf{S} = \mathbf{H}\mathbf{P}_{k+1|k}\mathbf{H}^T + \mathbf{R} \tag{15}$$

$$\mathbf{K} = \mathbf{P}_{k+1|k}\mathbf{H}^T\mathbf{S}^{-1} \tag{16}$$

Finally, the state vector and covariance matrix are updated:

$$\mathbf{x}_{k+1} = \mathbf{x}_{k+1|k} + \mathbf{K}\mathbf{y} \tag{17}$$

$$\mathbf{P}_{k+1} = (\mathbf{I}_{12} - \mathbf{K}\mathbf{H})\mathbf{P}_{k+1|k} \tag{18}$$

Through this iterative process, the Kalman Filter effectively fuses actual measurements with the imposed constraints, refining the estimates of the two points' positions and directions. The direction consistency constraint ensures that the directional vectors of the two points remain aligned, while the relative distance constraint maintains a stable spatial relationship between them. Proper tuning of the noise covariance matrices **Q** and **R** is essential to balance the influence of the process dynamics and the strength of the constraints on the state estimation.

5. The seed generation of railway-tracks

We construct the rough 3D line based on the dense point cloud. Given the end point **p** of a 2D line segment, we randomly sample three space points around **p** to construct the plane π, and we cast a ray **r** passing through **p** from the camera center. Then, the 3D point candidate $\mathrm{P} \in R^{3\times1}$ for **p** can be obtained by **r**-to-π intersection, and the candidate 3D line L can be represented by the two 3D point:

$$L = \left\{ inter\left(\pi, \mathbf{r}_1\right), inter\left(\pi, \mathbf{r}_2\right) \right\}, \tag{19}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the rays of the two endpoints, and *inter* calculates the ray-to-plane intersection. After random sampling of n times, we obtain a set of candidate 3D lines $\{L\}_{i=1}^{n}$ and use the LMEDS algorithm, which does not require an inlier threshold, to confirm the best 3D line for a 2D line:

$$L^* = \arg\min_{L_i} \text{median}\left\{ d_{i1}, d_{i2}, \ldots, d_{in} \right\}, \tag{20}$$

where d_{ij} is the distance between L_i and L_j.

We group two 3D lines as a RT pair based on their angle $\theta_{i,j}$, overlap $o_{i,j}$, and projection distance $d_{i,j}$:

$$\left\{ RT = \left(L_i, L_j\right) \mid \theta_{i,j} < t_\theta, o_{i,j} > t_o, d_{i,j} \in I \right\}, \tag{21}$$

$\theta_{i,j}$ and $o_{i,j}$ are easy to choose, e.g., 5° and 60%, because the RT pair is parallel and highly overlapped; while the interval I needs the rough width ω between the two RT, which can be acquired from construction standards or point clouds. We recommend setting $I = \left[2/3\omega, 4/3\omega\right]$ that uses one-third of ω as the margin of error. Because a 3D line may satisfy Eq. (21) with many others, the greedy algorithm is used to assign the candidate pair, which uses the sum of the overlap rate as the maximum score.

We sort the RT based on their scores of the geometry alignment and select the top 10% RT and use contextual information to further validate the RT pair. In detail, if the RT's central line is within 1° and t projection distance with another RT, its score is increased by $\mathcal{N}\left(\mu, (t/3)^2\right)$. we use the global average pooling layer in ResNet50 to describe the feature of the RT pair. Because it has been trained on massive amounts of data and can capture texture information for classification in the absence of labels. Also, we re-transform the image blocks to reduce the ambiguity caused by scale and rotation. After extraction of RT features, we use DBSCAN to group them with the cosine distance, and retain the group with the highest number as the seeds of RT.

References