**IMGTR: Image-triangle based Multi-view 3D Reconstruction for Urban Scenes**

Zhihua Hu ¹, Yaolin Hou ¹, Pengjie Tao ¹ and Jie Shan ²,*

1 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; huzhihua@whu.edu.cn; houyaolin@whu.edu.cn; pjtao@whu.edu.cn
2 Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA
*Correspondence: jshan@purdue.edu

**Abstract:** Multi-view depth map reconstruction is a popular approach to generate 3D information with good flexibility and scalability. However, texture-weak regions and repeated textures challenge these methods in urban scenes. To address this need, this paper proposes an image-triangle based multi-view 3D reconstruction (IMGTR) method. Starting from constructing a density-adaptive image triangulation for each image, the main procedure is to determine the corresponding object plane for each image triangle under an objective function consisting of the image similarity measure, the smoothness constraint and the continuity constraint on the edges between adjacent triangles. Qualitative and quantitative experiments show that the proposed method can reconstruct complex urban structures more accurate and achieve higher fidelity in planar urban structures than some recently released multi-view stereo algorithms. Specifically, for the Vaihingen dataset IMGTR achieves an average improvement of 4.27% compared to PMVD, 50.94% to SURE and 54.76% to COLMAP in position accuracy.

**Key words:** triangulation, multi-view reconstruction, depth map, urban scenes, continuity constraint

1. **Introduction**

3D reconstruction of urban scenes is a major task for photogrammetry and computer vision. Image-based 3D reconstruction could recover the scene from multiple images taken from different views. Multi-view depth map reconstruction aims to compute the distance from the corresponding object location for each pixel to the principle plane of the camera (Furukawa and Ponce, 2010; Shen and Hu, 2014). It can be adaptive to large-scale scenes due to its flexibility and scalability compared with other image-based
3D reconstruction methods, such as voxel-based (Faugeras and Keriven, 1998; Heise et al., 2015), deformable polygonal-meshes-based (Esteban and Schmitt, 2004; Furukawa and Ponce, 2009) and patch-based methods (Habbecke and Kobbelt, 2006).

However, there are still many challenges in achieving accurate multi-view depth map reconstruction for urban scenes due to the existence of texture-poor regions and repeated textures. Traditional methods regard a pixel as the basic unit to reconstruct depth value under the assumption of smooth geometry (Campbell et al., 2008; Shan et al., 2014). The smooth assumption of geometry is often imposed on the four-neighbor or eight-neighbor of a pixel to improve the robustness of the reconstruction and depress the noise. However, the existence of noisy, clutter and front-parallel bias in the reconstruction results limits its applications since not all scenes meet such smooth assumption. The introduction of image segmentation might handle this problem by imposing the smooth constraint over image segmentation. However, traditional image segmentation methods could hardly balance the segmentation scale and reconstruction details, which means that they could not take into account the abundant details in the urban scenes. Meanwhile, the existence of strong structural regularities, including flat, sharp corners and line segments (Furukawa et al., 2009) also challenges this kind of methods. Moreover, the boundary of image segments may not be at the edge of the object in urban scenes, which makes the reconstruction results fail to represent the geometric shape of the object appropriately. Furthermore, these methods use the segments as the basic unit to conduct reconstruction. As such, the final results would still be represented by pixels, yielding a lot of redundancy.

In this paper, we propose an image-triangle based multi-view 3D reconstruction method for urban scenes (IMGTR). Firstly, a density-adaptive image triangulation is constructed using line segments and corner points obtained through feature extraction, i.e. LSD (Line Segment Detector; Grompone et al., 2012) and Harris (Harris and Stephens, 1988). The reconstruction is then transformed to determine the corresponding object planes for each image triangle. To this end, IMGTR forms an energy function based on an image similarity measure of the image triangles, smoothness constraints and continuity constraints on the common edge of the adjacent triangles in the object space. Finally, the corresponding plane for each image triangle can be determined through minimizing the energy function with an optimization framework similar to the Expectation Maximization (EM) (Dempster et al., 1977) optimization. Instead
of building a global 3D-mesh model, IMGTR reconstructs the depth map and normal map for each image since it is hard to precisely detect planes in the object space without extra information. IMGTR is evaluated qualitatively and quantitatively on several datasets with different scales and resolutions. The results demonstrate that IMGTR can preserve the planar urban structures and achieve higher reconstruction accuracy than the recently released algorithms, including SURE (Rothermel and Wenzel, 2012), COLMAP (Schönberger et al., 2016) and PMVD (Hou et al., 2018). In addition, much less points are needed for IMGTR to represent the same scene compared to pixel based dense matching method since a density-adaptive image triangulation is constructed with line segments and corner points.

2. Related Works

There are many remarkable works about image-based 3D reconstruction in the past few decades. By using a wide range of scenes from large, shared, multi-user photo collections available on the Internet, Goesele et al. (2007) effectively handled the large differences in illumination or scale present in these images through global and local view selection to assure high quality reconstructions of landmarks in urban scenes. A system designed to maximize parallel computing was proposed to reconstruct 3D geometry from large, unorganized image collections and achieved building a city in less than a day (Agarwal et al., 2011).

Image-based 3D reconstruction for Lambertian surface with abundant textures has achieved considerable development and could obtain relatively ideal reconstruction results (Furukawa and Hernández, 2015). To handle more diverse data, many methods converted the 3D reconstruction into an energy optimization problem. Through minimizing the energy function often composited by a data term and regularization term, these methods could improve the reconstruction robustness and performance (Campbell et al., 2008; Kolmogorov & Zabih, 2001; Li et al., 2015). Smoothness assumption is the most widely used regularization constraints, which encourages that the points in a local area change smoothly (Shan et al., 2014). It should be noted that the generic formulation of such optimization framework makes it possible to employ specialized constraints beyond the common smoothness on final reconstruction results.

Using structural features as priors to impose constraints during 3D reconstruction procedure was
shown effective to take care of the less texture or repeating texture area (Furukawa and Hernández, 2015; Xu and Tao, 2020). Urban scenes often consist of regular shapes including flat objects, sharp corners and lines. These distinct geometric characteristics provide useful information to formulate effective regularization constraints to improve the reconstruction results (Hou et al., 2018). According to the space in which the constraint is imposed, these Multi-View Stereo (MVS; Furukawa and Hernández, 2015) methods considering the structural features can be roughly divided into two categories: methods based on structural features either in object space or in image space.

The object space structural feature based MVS methods generally extract different geometric primitives in the initial dense point cloud and apply them in the object space to achieve complete and accurate 3D information. Typically, a relatively accurate initial point cloud or mesh is required as a prior. Lafarge et al. (2013) proposed a hybrid multi-view reconstruction which models urban scenes as a combination of meshes and geometric primitives to simultaneous represent irregular elements and regular structures at the same time. Firstly, the initial surface is effectively segmented by multi-label Markov Random Field (MRF) optimization. Then, the geometric primitives and mesh components are extracted simultaneously on the segmentation results. In response to the noise present in the initial point cloud, a method based on dividing the horizontal structures into slices have been proposed, which can reconstruct walls and flat roofs very well (Holzmann et al., 2016). However, in the slanted plane, this method may cause serious front-parallel. Based on this method, Holzmann et al. (2017) further proposed a three-dimensional reconstruction method with mixed voxel reconstruction and plane fitting, which was shown to be robust to noise. This method first constructs a tetragonal pyramid by using the Delaunay triangulation. 3D scene representation is then obtained by optimizing the internal or external attributes of the voxels.

The image space structural feature based methods typically apply the structural features in the image space with a standard MRF formulation to achieve multi-view depth map reconstruction through global optimization. For example, the Manhattan-World assumption, i.e., the major components of a building consist of axis-aligned planes, is widely used in urban scenes reconstruction (Coughlan & Yuille, 1999; Li et al., 2016). Furukawa et al. (2009) proposed a novel method for Manhattan-World and achieved good performance in less texture scenes. This method extracts dominant plane directions from the initial point
cloud and generates plane hypotheses, then recovers pixel-wise depth for each input image (Furukawa et al., 2009). Sinha et al. (2009) acquired a discrete set of 3D plane candidates from a sparse point cloud and 3D line segments, and generated piecewise planar results under the constraints imposed by vanishing directions, 3D lines, and other high-level vision features. A method designed for speed and efficiency generated plane hypotheses by an iterative clustering step and assigned each pixel an optimal hypothesis through semi-global matching (Sinha et al., 2014).

Due to the simultaneous existence of piecewise planar objects and non-planar objects, Gallup et al. (2010) proposed to segment images and apply different structural constraints to different regions. However, its performance for details in the scene is not satisfactory. With an over-segmentation of the image, some methods set the segmentation as a basic unit and determined the optimal local plane for each unit. Zebedin et al. (2008) proposed to use the projection of the 3D line segments to obtain a line-based segmentation of the building and set aerial imagery as input data to generate corresponding elevation data. Duan and Lafarge (2016) introduced a method to generate compact and semantic-aware city models from satellite images. This method could improve the robustness to occlusion and low image quality by combining classification and elevation estimation. However, these two methods could not handle the façade effectively. An edge-aware surface mesh reconstruction method for urban scenes was then introduced (Bódis-Szomorú et al., 2017) that uses the image edge to construct 2D base mesh by constrained Delaunay triangulation and generates compact, edge-preserving meshes from sparse cloud point or height map. However, its performance highly depends on the quality of the input sparse cloud point or height map. Hou et al. (2018) proposed a planarity constrained multi-view depth map reconstruction method which could preserve the planarity of piecewise flat structures and the linearity of discontinuous edges. Nevertheless, other geometric characteristics beyond planarity were not considered and there may be discontinuous cases in the flat areas.

Based on the related works, we propose a multi-view three-dimension reconstruction method with the use of geometric primitives, in particular, line segments and corner points extracted in the images. Compared to our earlier work (Hou et al., 2018), the urban scenes can be represented more efficiently since less planes are used to describe the same scene due to the introduction of line segment. Besides, the proposed method can avoid noises or over-smooth in the discontinuous areas by carefully handling the
depth-fracture. Unlike the previous edge-aware 3D reconstruction work (Bódis-Szomorú et al., 2017) which needs extra sparse cloud point or height map, the proposed method utilized the extended PatchMatch (Hou et al., 2018) to achieve acceptable results even with random initialization.

3. Method

This section will describe the proposed method in detail. As shown in the Figure 1, the input of IMGTR is an image set $J$ of $N$ images as well as their camera parameters. Firstly, line segments and corner points are obtained by feature extraction (LSD and Harris in our case) and applied to construct triangulation on each image $I_i$. Then, an energy function is built on a 2D triangulation in the images with the image similarity constraint, smoothness constraint and continuity constraint. Next, the local planes of the triangles in the 2D triangulation are initialized with feature matching (SIFT matching in our case) and random values. Finally, the depth map of each image is reconstructed through solving an energy function with an iterative two-step optimization scheme similar to the EM optimization (Dempster et al., 1977), in which the “E” step determines the depth-fracture and the “M” step assigns optimal planes for each triangle in the images.

Figure 1. Flowchart of the proposed image-triangle based multi-view 3D reconstruction (IMGTR) method. The normal map is rendered with the normal vector of the surface and all the normal maps in the following figures use the same color code.
3.1 Image triangulation generation

A triangle of the 2D triangulation in the images should not overlap two different planes in the object space. To this end, we extract both corner points and line segments in the images. For corner points extraction, the classic Harris (Harris and Stephens, 1988) is used. To ensure enough corner points, the threshold of the classic Harris corner response function is set relatively low. Then, in order to make the corner points well-distributed, the detected corner points are clustered according to their distances. For each cluster, eight (8) corner points with maximum possibility to be corner points in a 5 by 5 pixels window are selected. As for line segments, the segment boundary generated by classic segmentation method might not be aligned with image edges, as these methods only consider the intensity and compactness. The LSD (Grompone et al., 2012) is chosen in IMGTR since the line segments extracted with LSD can better describe the edges of the objects in the urban scene while the default parameters in OpenCV (Bradski, 2000) is used. However, the line segments extracted with LSD cannot be used directly to construct 2D triangulation in the images, since there is no topology, e.g. intersection of line segments. As such, cleaning the extracted line segments is needed before 2D triangulation.

As shown in Figure 2, the cleaning consists of merging, extension, and snapping. To be specific, if the angle of two line segments is less than, e.g., 10 degrees and their distance is less than 2 pixels, these two line segments will be merged. For example, the blue and red line segments in Figure 2a are merged into one line segment. When the end points of the two line segments are within 12 pixels, they are extended to intersect with each other as shown in Figure 2b. Similarly, intersections are snapped for overshooting and undershooting cases as shown in Figure 2c and 2d.

![Figure 2](image.png)

Figure 2. Cleaning of the extracted line segments (top) to assure correct topology (bottom)
After line segment cleaning, the Constrained Delaunay triangulation (CDT) (CGAL, 2018) is used to generate image triangulation which can guarantee the line segments are the edges of the triangles. When constructing the 2D triangulation, the triangles should meet the following two conditions: the number of pixels in the triangle $N_e > \tau_ne$; any angle of the triangle $|\cos(\theta_t)| > \tau_{na}$. Firstly, we use the line segments as constraints to construct the image triangulation. Then, corner points are added to the image triangulation based on the two above conditions. The line segments or corners near a triangle which does not meet these conditions should be deleted from constraints.

Figure 3. Image triangulation result (left): (a) result with line segments and corner points in the red area; (b) result with only line segments in the red area, and (c) result with only corner points in the red area.

Figure 3 shows the image triangulation with the constraints of line segments and corner points. It shows that the triangulation with the constraints of line segments alone can be aligned with image edges, while the triangulation with the constraints of corner points alone can describe the details very well, especially in the areas with the dormer window and chimney. Our method can therefore combine the
advantages of both line segments and corner points and achieve a density-adaptive image triangulation. This means there are many triangles in the rich texture area, while the number of triangles in the weak-texture region is fewer.

Comparing to current MVS methods that only use line segmentation as constraints and cannot describe the geometric details well, IMGTR can generate density-adaptive image triangulation with the corner points and line segments and describe the geometric details of the urban scenes well. It can balance the number of segments and the size of the segments.

3.2 Energy function

Next, IMGTR formulates the image-based reconstruction into an energy minimization procedure. The main notations used in the energy function with their relationships are illustrated in Figure 4. For the input image set $\mathcal{I}$, each image is in turn treated as a reference image denoted as $I_r$, and the rest of the images in $\mathcal{I}$ are denoted as target image set $\mathcal{I}' = \{ I'_i \} = \mathcal{I} \setminus I_r$. After image triangulation, we obtain the $\mathcal{T}_r = \{ T_j \}$ consisting of triangle $T_j$ with the vertexes $v$ and edges $e$, where the integer pixel set is $\mathcal{X}_j = \{ x_n \}$. Then, for each triangle $T_j$, IMGTR assigns an optimal local plane $\ell_j = [n^T, D]$ described by normal $n$ and another parameter $D$ which satisfies the equation $n^T x_{T_j} + D = 0$ where $x_{T_j}$ is the corresponding object coordinates of the pixels in triangle $T_j$. IMGTR uses image similarity constraint, smoothness constraint and continuity constraint to form an energy function for optimization.

![Figure 4. Notations used in the IMGTR method and their relationships.](image-url)
Next, we describe the details of the three constraints in the energy function, i.e. the image similarity constraint, the smoothness constraint and the continuity constraint. The image similarity constraint is the “data term” in our energy function. It measures the similarity of each 2D image triangle with the color and texture cues. However, by only using image similarity constraint, i.e. treated every 2D triangle independently, it will result in non-planar triangles and discrete edges among adjacent triangles in object space. Thus, a smoothness constraint is built to constrain that triangles sharing the same edges should be in the same local plane. A continuity constraint is built to enforce the continuity of common edges of the adjacent triangles in object space.

**Image similarity constraint.** The image similarity measures how well a pixel $\mathbf{x}$ in the reference image is matched with its corresponding pixels in the other images. In IMGTR, the image similarity constraint is defined as:

$$E_{\text{similarity}}(\mathbf{e}) = \sum_{T_j \in T_r} C(T_j, \mathbf{e}_j) = \sum_{T_j \in T_r} \frac{1}{|X_j|} \sum_{x_n \in X_j} \varphi(x_n, \mathbf{e}_j)$$

(1)

where $C(T_j, \mathbf{e}_j)$ is a matching cost of the triangle $T_j$ which is a function of the local plane $\mathbf{e}_j$. $|X_j|$ is the number of pixels in $X_j$ and $\varphi(x_n, \mathbf{e}_j)$ is the matching cost of the pixel with respect to $\mathbf{e}_j$. For all of the target images, we conduct pixel-wise view selection to determine the effective target image set $I_{ne}$ for pixel $x_n$ and compute its matching cost using the following average:

$$\varphi(x_n, \mathbf{e}_j) = \frac{1}{N_e} \sum_{m \in I_{ne}} C(x_n, x_{\pi, m}, \mathbf{e}_j)$$

(2)

where $N_e$ is the number of these images in $I_{ne}$ and $C$ is the matching cost of $\mathbf{x}$ with the corresponding pixel $x_{\pi, m}$ on target image $I'_m$. As the optimization process is iterative, we could use 3D information obtained in last iteration to select effective target set $I_{ne}$ for pixel $x_n$. IMGTR calculates a score of each target image and selects the top 50% images as the final set. The score function is defined as:

$$S_m = \rho_t(x_n, m)\rho_r(x_n, m)\rho_v(x_n)$$

(3)

The triangulation term $\rho_t(x_n, m) = n(C_r, X) \cdot n(C_m, X)$ ensures that the viewpoint change between the reference image and the target image is appropriate. $X$ is the object point corresponding to $x_n$, $n(C_r, X)$ and $n(C_m, X)$ are the normal vectors from $X$ to the camera center $C_r$ of the reference image and the camera center $C_m$ of the target image respectively. The scale term $\rho_r(x_n) = (X - C_m)^T n_j / \|X - C_m\|$ and
ensures that the angle of the projective ray and the surface normal vector \( \mathbf{n}_j \) is small.

Considering the color and texture cues, the matching cost of single pixel \( C(\mathbf{x}, \mathbf{x}_\pi, \mathbf{e}_j) \) is the defined as:

\[
C(\mathbf{x}, \mathbf{x}_\pi, \mathbf{e}_j) = \beta \cdot \min(C_{\text{census}}(\mathbf{x}, \mathbf{x}_\pi, \mathbf{e}_j), \tau_t) + (1 - \beta) \cdot \min(\|I_{rg}(\mathbf{x}) - I'_{mg}(\mathbf{x}_\pi, \mathbf{e}_j)\|, \tau_c)
\]

where \( \beta \) is a ratio \( \beta \in [0,1] \) used to balance the effort of the two terms, \( I_{rg}(\mathbf{x}) \) and \( I'_{mg}(\mathbf{x}_\pi, \mathbf{e}_j) \) are the gray values of pixel \( \mathbf{x} \) in reference image \( I_r \) and pixel \( \mathbf{x}_\pi \) in the target image \( I'_m \). \( \tau_t \) and \( \tau_c \) are truncation values. Census-based term \( C_{\text{census}} \) is the Hamming distance of two bit strings that stand for pixel \( \mathbf{x} \) and its correspondence \( \mathbf{x}_\pi \), which is calculated in the same way as (Hou et al., 2018). With the \( \mathbf{e}_j = [\mathbf{n}, D] \), we could obtain the corresponding pixel \( \mathbf{x}_\pi(x, y) \) in a target image \( I'_m \) for the pixel \( \mathbf{x}(x, y) \) in the reference image under the planar geometry model (Galliani et al., 2015; Hou et al., 2018; Schönberger et al., 2016). When the camera coordinate system of a reference image is set as the world coordinate system, depth value \( d \) is equal to \( Z \) value of that object point, which can be calculated by local plane and camera parameters:

\[
d = Z = -Df_x f_y / (f_y (y - v)) \cdot \mathbf{n}
\]

where \((u, v)\) is the image coordinates of the principal point and \( f_x, f_y \) represent the focal lengths of the camera. The projection relationship of the reference image and target image could be described as:

\[
\mathbf{x}_\pi = \mathbf{H}_\pi \mathbf{x}; \quad \mathbf{H}_\pi = \mathbf{K}_m \left( \mathbf{R}_m - \frac{1}{D} \mathbf{T}_m \mathbf{n}^T \right) \mathbf{K}^{-1}
\]

where \( \mathbf{K}_m, \mathbf{R}_m, \) and \( \mathbf{T}_m \) are the camera parameters of target image \( I'_m \), \( \mathbf{K} \) is the intrinsic parameter matrix of the reference image.

**Smoothness constraint.** The smoothness constraint is to encourage that the triangles with common edges have the same local plane, which could reduce the influence of noise and less texture on reconstruction and improve the reliability of the reconstruction. The smoothness constraint is defined as:

\[
E_{\text{smooth}}(\mathbf{e}) = \sum_{T_j \in \mathcal{T}} \sum_{u \in N_j} w_{uj} \cdot (\mathbf{e}_u - \mathbf{e}_j)^T W_{nd}(\mathbf{e}_u - \mathbf{e}_j)
\]

where \( \mathcal{N}_j \) is a set of triangles with a common edge to \( T_j \), \( \mathbf{e}_u \) and \( \mathbf{e}_j \) are local plane of \( T_u \) and \( T_j \) respectively. \( w_{uj} = \exp \left( -\|c_u - c_j\|_2 / \sigma \right) \) is a weight function used to avoid over-smoothing at some edge, where \( c_u \) and \( c_j \) are the average color intensity of pixels in \( T_u \) and \( T_j \), \( \sigma \) is a constant value. \( W_{nd} = \text{diag}(R_d, R_d, R_d, 1) \) is a diagonal matrix to balance the effect of the normal vector \( \mathbf{n} \) and the
parameter $D$ of the local plane. $R_d$ denotes the range of the depth computed by the sparse point cloud generated with feature (e.g. SIFT) matching, i.e. $R_d$ is calculated by subtracting the smallest depth from the biggest depth.

**Continuity constraint.** The range of image-based 3D reconstruction result should be continuous in most locations. Traditional MVS methods might only get reconstruction result at an accuracy of pixels, which would cause the appearance of front-parallel bias. PatchMatch tries to find the optimal 3D plane for every pixel and makes it possible to achieve sub-pixel level accuracy (Bleyer et al., 2011). However, it is difficult to impose the continuity constraint on the discrete points since the image is the discrete representation of object information. To address this limitation, we determine an optimal local plane for each triangle in IMGTR. As such, the 3D information is continuous within the triangle. If there were no continuity constraint, the adjacent triangles would exhibit obvious drift at common edge, whereas the common edge can be better reconstructed with continuity constraint as shown in Figure 5.

![Figure 5. Reconstruction result on image triangulation without and with continuity constraint: (a) image triangulation; (b, c) reconstruction without or with continuity constraint.](image)

To enforce the continuity of the adjacent triangles at common edge, IMGTR imposes continuity constraint as follows:

$$E_{continuity}(\mathbf{e}) = \sum_{e_n \in \mathcal{E}} \tau \cdot (1 - \alpha) \cdot \left( |d(v_p, \mathbf{e}_j) - d(v_p, \mathbf{e}_u)| + |d(v_q, \mathbf{e}_j) - d(v_q, \mathbf{e}_u)| \right)$$

(8)

where $e_n$ is the common edge of the triangle $T_j$ with local plane $\mathbf{e}_j$ and the triangle $T_u$ with local plane $\mathbf{e}_u$, $\mathcal{E}$ is the set with all common edges, $v_p$ and $v_q$ are the end points of edge $e_n$, $d(x, \mathbf{e})$ is a depth value function described by Equation 5. $\tau$ is the weight of the continuity constraint to balance the energy terms. $\alpha$ indicates whether the edge $e_n$ is a depth fracture. If $e_n$ is a depth fracture, it means that the adjacent triangles in images are actually not adjacent in object
space, e.g. the edges of building roof and ground can be the same edge in the image space, while in object space, they should be two different edges. There should be no continuity constraint on the depth fracture, thus $\alpha$ is set to be 1 when $e_n$ is a depth fracture. IMGTR introduces a depth-fracture determination method based on graph cut (Boykov et al., 2001; Kolmogorov, 2004).

**The final energy function** in IMGTR is the summation of the above three terms (Equation 1, 7 and 8):

$$E(\ell) = E_{\text{similarity}}(\ell) + E_{\text{smooth}}(\ell) + E_{\text{continuity}}(\ell)$$  \hspace{1cm} (9)

By setting local plane $\ell_x = [n^T, D]$ as the optimization variable, IMGTR assigns an optimal local plane for each triangle through the energy minimization process.

### 3.3 Optimization Solution

The solution of the optimization problem (Equation 9) is of a challenge since most of the variables are continuous and coupled. Inspired by (Zhang and Li, 2015), we introduce a two-step approach to optimize the energy function iteratively, where the first step is to update binary variable $\alpha$ by graph cut and the second step is to assign the optimal local plane to each triangle by quadratic relaxation. Table 1 gives the pseudo-code of the optimization procedure.

**Table 1. Pseudo-code of the optimization procedure in IMGTR.**

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>images $I$, camera parameters $C$, 2D triangles and feature matching</td>
<td>depth maps $Z_{\text{map}}$, normal maps $N_{\text{map}}$</td>
</tr>
</tbody>
</table>

1 for each input image $I_i$ do  
2 Initialization of local planes for the whole optimization  
3 for iteration $t_1 = 0$ to $t_{1\text{end}}$ do  
4 for each input image $I_i$ do  
5 depth-fracture determination through optimizing the Eq. 10 E-step  
6 for iteration $t_2 = 0$ to $t_{2\text{end}}$ do  
7 for each triangle in $I_r$ do  
8 obtain the optimal local planes through random search, internal propagation, outlier elimination and external propagation (Hou et al., 2018) M-step  
9 optimizing Eq. 17 with Levenberg–Marquardt algorithm  
10 Return $Z_{\text{map}}$ and $N_{\text{map}}$

**Initialization.** To solve the energy function, we firstly initialize the local plane with feature matching (SIFT) and random values. For the triangles with enough feature points, plane fitting is used to calculate the initial local plane. For the rest triangles, the local plane is initialized with random values.

**E-step.** For a local plane $\ell$, the first step is to determine whether an edge is a depth fracture through
the optimization of Equation 10 with graph cut. The energy function is constructed with image intensity and depth value in current iteration as follows:

\[
E = \sum_{e_n \in \mathcal{E}} E_{data}(e_n, \alpha_n) + \sum_{e_n \in \mathcal{E}} \sum_{e_m \in \mathcal{E}} E_{\alpha}(e_n, \alpha_n, e_m, \alpha_m)
\]

(10)

where \(\mathcal{E}_n\) is a set of edges with one end point of \(e_n\). The data term \(E_{data}(e_n, \alpha_n)\) is defined as Equation 11:

\[
E_{data} = \log(\hat{g}_e) \times Diff(e, \ell_j, \ell_u)
\]

(11)

where \(\hat{g}_e\) is the edge gradient, which is the average of gradient magnitude along the edge \(e_n\); \(Diff(e_n, \ell_j, \ell_u)\) is the average difference of depth values of end points \(v_p\) and \(v_q\) calculated by the local planes \(\ell_j\) and \(\ell_u\) corresponding to \(T_j\) and \(T_u\), which are adjacent triangles with common edge \(e_n\).

The smooth term \(E_s\) encourage the edges share the same end point tend to have the same fracture state. Similar to (Zhang et al., 2015), we introduce visual complexity levels (indicate pixels in the images are in a same visual levels or not) as an auxiliary variable defined as:

\[
k(e) = \arg\max_j \{|\hat{g}_e^l - \hat{g}_e| < \sigma, \forall l \leq j\}
\]

(12)

where \(\{\hat{g}_e^l, l = 1, 3, 5, 7, 9\}\) is a sequence of edge gradient, \(\hat{g}_e^l\) is calculated with Gaussian-blurred images using kernel size \(l \times l\). If the end points of edges are of similar visual complexities, they should have the same fracture state. Thus, the smooth term is defined as:

\[
E_s(e_n, \alpha_n, e_m, \alpha_m) = (V_{max} - |k(e_n) - k(e_m)|) \times |\alpha_n - \alpha_m|
\]

(13)

where \(V_{max}\) is the largest visual complexity level of the scene and is set to 5 in our case. Graph cut (Kolmogorov, 2004) is used to conduct binary optimization.

**M-step.** Keeping the binary variable \(\alpha\) unchanged, the second step is to acquire the optimal local plane for each triangle by minimizing Equation 9. Inspired by (Zhang & Li, 2015; Zhang et al., 2014), we adopt the quadratic splitting technique to relax the energy function to solve this problem. By introducing an auxiliary variable \(\tilde{\ell}\) this minimization problem is converted to two sub-problems, which in turn can be solved as described below:

\[
E_{retexed}(\ell, \tilde{\ell}) = E_{similarity}(\tilde{\ell}) + \frac{\theta}{2} E_{couple}(\ell, \tilde{\ell}) + E_{smooth}(\ell) + E_{continuity}(\ell)
\]

(14)

where \(E_{couple}(\ell, \tilde{\ell})\) is a coupling term used to relate \(\ell\) and \(\tilde{\ell}\), defined as follows:

\[
E_{couple}(\ell, \tilde{\ell}) = (\ell - \tilde{\ell})^T W (\ell - \tilde{\ell})
\]

(15)

where \(W = \text{diag}(R_d, R_d, R_d, 1)\) is a diagonal matrix to balance the contributions of the normal vector.
and the other parameter of $D$ the local plane. In Equation 14, $\theta$ is a control factor. When $\theta$ tends to be infinity, the energy function (Equation 9) is equal to the energy function (Equation 14) after relaxation. In the optimization process, we use the same setting strategy as (Zhang and Li, 2015) which increases $\theta$ from 0 to 100.

IMGTR updates $\mathbf{\tilde{e}}$ and $\mathbf{e}$ iteratively. First, by fixing $\mathbf{e}$, we update $\mathbf{\tilde{e}}$ by minimization of following energy function:

$$
\min_{\mathbf{\tilde{e}}} E_{similarity}(\mathbf{\tilde{e}}) + \frac{\theta}{2} E_{couple}(\mathbf{\tilde{e}}, \mathbf{\tilde{e}})
$$

(16)

IMGTR adopts the extended PatchMatch method (Hou et al., 2018) to solve this problem. Then, by fixing $\mathbf{\tilde{e}}$, the update of $\mathbf{e}$ reduces to:

$$
\min_{\mathbf{e}} \frac{\theta}{2} E_{couple}(\mathbf{e}, \mathbf{\tilde{e}}) + E_{smooth}(\mathbf{e}) + E_{continuity}(\mathbf{e})
$$

(17)

This is a nonlinear least squares problem, for which IMGTR adopts the Levenberg–Marquardt algorithm (Marquardt, 1963). By now, we can minimize the energy function (Equation 9) constructed by image similarity constraint, object smoothness constraint and object continuity constraint iteratively, and obtain the optimal local object plane corresponding to each triangle in the image triangulation.

4. Experiments

This section will evaluate image-triangle based multi-view 3D reconstruction method (IMGTR). The experiments use aerial images and unmanned aerial vehicle (UAV) oblique images as test data for urban scenes. The experiments are performed on a PC with Intel i7-4790 CPU (3.60Hz) and 32GB RAM; and the implementation is conducted with C++. The parameter settings in IMGTR across all the datasets are listed in Table 2.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$</td>
<td>factor used to control the ratio of the smoothness term</td>
<td>15</td>
</tr>
<tr>
<td>$\tau$</td>
<td>weight of the continuity constraint</td>
<td>0.5</td>
</tr>
<tr>
<td>$\sigma_g$</td>
<td>threshold of the gradient difference in the calculation of visual complexity level</td>
<td>3.5</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>Maximum visual complexity level</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma_N$</td>
<td>the minimum number of integer points in the triangle</td>
<td>6</td>
</tr>
<tr>
<td>$\sigma_{\theta_t}$</td>
<td>threshold of the cosine of the largest angle of the triangle</td>
<td>0.999</td>
</tr>
<tr>
<td>$\beta$</td>
<td>weight of Census-based matching cost</td>
<td>0.9</td>
</tr>
</tbody>
</table>
4.1 Qualitative evaluation

Two oblique image datasets are used for the qualitative evaluation. The first one is the Dortmund data provided by International Society for Photogrammetry and Remote Sensing and EuroSDR (Nex et al., 2015). This data was collected in May 2014 in the city center of Dortmund, Germany, using AeroWest's PentaCam airborne oblique image system with five cameras. The ground spacing distance (GSD) is 10 cm in nadir images and varies from 8 to 12 cm in the oblique views. The along flight and across flight overlap are 75% and 80%, respectively. As a result of bundle adjustment, the back-projection error is 0.122 pixels. In this paper, a subset of 17 images shown in Figure 6 from the nadir view camera and a side view camera is selected for the experiment. This area is a typical representation of urban scenes with buildings of complex structures.

![Dortmund data with 17 images in total in which the image size for nadir view (a) is 1601 × 2191 pixels and the image size of side view (b) is 2480 × 1630 pixels.](image)

Figure 6. Dortmund data with 17 images in total in which the image size for nadir view (a) is 1601 × 2191 pixels and the image size of side view (b) is 2480 × 1630 pixels.

Figure 7 shows the normal vector reconstruction results (same legend as Figure 1) for the image in Figure 6a. Compared to the initial normal vector map from feature matching, IMGTR can reconstruct the scene with much more accurate normal vectors which express the geometry of the urban scenes well, especially in the building roofs. It indicates that IMGTR is robust to initial noisy normal map.
Figure 7. Reconstruction results of Dortmund data: the normal vector map of the image in Figure 6a; the left is the initial value from feature matching while the right is the final result from IMGTR.

Figure 8 shows the qualitative evaluation results of the Dortmund data, including the elevation and normal vector maps corresponding to the two images a) and b) in Figure 6. The elevation of each pixel is calculated based on the optimal object plane corresponding to the triangle it is located at. Through this figure, we could find that the scene geometry is satisfactorily reconstructed, where the outline is sharp and the edges are effectively preserved. In particular, the edge in the depth discontinuous areas is well recovered without over smoothing. Moreover, there is no obvious discontinuity in the flat areas.
Figure 8. Reconstruction results of Dortmund data: (a) and (b) are the corresponding results of the image in Figure 6a; (c) and (d) are the corresponding results of the image in Figure 6b; the left is the corresponding elevation map, while the right is the normal vector.

Another dataset used for qualitative experimental analysis is the UAV oblique dataset over northwestern Tongchuan City in Shanxi Province, China, collected by a platform with five SONY ILCE-7R cameras (left, right, nadir, front, and back). The nadir ground sampling distance is about 6 cm. As shown in Figure 9, for this study we select a subset of seven images as a representative of semi-urban scenes with residential buildings and low level houses.

Figure 9. The unmanned aerial vehicle (UAV) oblique dataset in Tongchuan
Figure 10 shows the reconstruction result in form of surface mesh. It is generated by projecting the image triangulation of the selected reference image to the object space with its camera parameters. As shown in the figure, the image-triangle based multi-view 3D reconstruction method could describe the geometric shape of the object well and achieve the effective expression of scenes through triangulation of varying densities. That means for areas with rich texture and complex structure, the triangles with a higher density are used to describe the detail, whereas for areas with simple structure, large size triangles are used to represent the surface. In other words, IMGTR can achieve a density-adaptive expression for the scene.

![Reconstructed triangle mesh without texture for the Tongchuan dataset.](image)

Figure 10. Reconstructed triangle mesh without texture for the Tongchuan dataset.

In order to show the results more convincingly, IMGTR is compared with other methods, including COLMAP which is regarded as the state of the art in open source projects, and PMVD (Hou et al., 2018) which is a planarity constrained multi-view depth map reconstruction method for urban scenes. The comparison results are shown in Figure 11, where (a), (b) and (c) are normal vector map generated by COLMAP, PMVD and IMGTR. It is shown that both IMGTR and PMVD can make better shape recovery for flat areas such as roofs. In contrast, the result of COLMAP has a lot of noise and clutter in planar areas. Figure 11(e), (f), and (g) are partial enlargements of the corresponding region in (d) for the three methods, respectively. Figure 10 shows that with the planarity constraint, IMGTR and PMVD are considerably superior in the description of planar areas. Compared with PMVD, IMGTR can avoid the appearance of noise point and obtain better performance in the discontinuous edge areas. The result of COLMAP has more details, but many of these details are false or exaggerated details, i.e., artefacts. Figure 11(g) shows the reconstruction results of a basketball court. It is obvious that the details at the edge areas are artefacts,
which is also a common problem in the traditional depth map reconstruction or dense matching algorithm. IMGTR can better describe the planar structures in the urban scenes and have less noise points than the recently released multi-view methods.

Figure 11. Normal vectors generated from the Tongchuan dataset: a), b), c) are the normal vector maps for a building respectively from COLMAP, PMVD and IMGTR, d)-g) show three corresponding image clips (d) and their normal vector maps from COLMAP (e), PMVD (f) and IMGTR (g).
4.2 Quantitative evaluation

At first, the Vaihingen dataset (Cramer, 2010) is used to quantitatively evaluate the IMGTR method. The dataset was provided by the German Association of Photogrammetry and Remote Sensing (DGPF) and consists of 20 pan-sharpened color infrared images with a ground sampling distance of 8cm (Cramer, 2010). As shown in Figure 12, three representative areas are selected for test. In addition to the images, the Vaihingen dataset also has airborne laser scanning (ALS) data with a median point density of 6.7 points/m², which could be regarded as ground truth for the quantitative evaluation. We use the integer-valued points in the triangle as the unit to conduct quantitative evaluation and calculated the depth value according to the optimal local plane of the triangle. For a pixel \( x \) with the corresponding object point \( X \) in the reference image, we determine a set \( P_x \) which consists of laser points whose XY-plane distance from \( X \) is less than a certain threshold. The threshold is fixed as 0.25m. The average elevation \( Z_g \) of the points in set \( P_x \) is used as the ground truth of pixel \( x \). The elevation RMSE \( \sigma_Z \) of \( Z_i \) of the checked pixel \( x \) is defined as follows:

\[
\sigma_Z = \sqrt{\frac{\sum_{i=1}^{N_e} (Z_i - Z_g)^2}{N_e}}
\]

(18)

where \( N_e \) denotes the number of check pixels. Table 2 shows the results of the quantitative evaluation. To avoid influence of large magnitude near the object boundaries, we adopt the same strategy with (Hou et al., 2018) by selecting the middle 95% of the elevation differences for our quantitative evaluation.

Figure 12. The three subsets with (a) 1777×2085 pixels, (b) 1499×1698 pixels and (c) 1537×2061 pixels in the Vaihingen dataset.

We also compare IMGTR with SURE, COLMAP and PMVD. As for SURE, a trial version of SURE
4.1 is used. Table 3 shows the qualitative results of the Vaihingen dataset. It should be noted that the same outlier elimination strategy as the one in (Hou et al., 2018) is adopted and the results shown below are evaluated after outlier elimination. It should be noted that there are no significant differences between IMGTR and our recent work PMVD (Hou et al., 2018) in terms of number of remaining pixels after outlier elimination, e.g. the number of remaining pixels in Area a is 2,949,616 and 2,771,575 for IMGTR and PMVD respectively. The average elevation accuracy of IMGTR reaches 0.371m RMSE. In these areas, IMGTR yields an average 4.27% improvement in position accuracy compared with PMVD, 50.94% compared with SURE and 54.76% compared with COLMAP.

**Table 3. Comparative quantitative evaluation of IMGTR on the Vaihingen dataset**

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma_Z$ (m)</th>
<th>$\sigma_Z$ (95%) (m)</th>
<th>Relative to altitude ($10^{-4}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Area a</strong> (1777×2085)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMGTR</td>
<td>1.195</td>
<td>0.431</td>
<td>4.8459</td>
</tr>
<tr>
<td>PMVD</td>
<td>1.205</td>
<td>0.445</td>
<td>5.0033</td>
</tr>
<tr>
<td>SURE</td>
<td>1.447</td>
<td>0.718</td>
<td>8.0469</td>
</tr>
<tr>
<td>COLMAP</td>
<td>1.562</td>
<td>0.800</td>
<td>8.9947</td>
</tr>
<tr>
<td><strong>Area b</strong> (1499×1698)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMGTR</td>
<td>1.297</td>
<td>0.364</td>
<td>4.0926</td>
</tr>
<tr>
<td>PMVD</td>
<td>1.412</td>
<td>0.375</td>
<td>4.2163</td>
</tr>
<tr>
<td>SURE</td>
<td>1.550</td>
<td>0.677</td>
<td>7.6073</td>
</tr>
<tr>
<td>COLMAP</td>
<td>1.767</td>
<td>0.711</td>
<td>7.9940</td>
</tr>
<tr>
<td><strong>Area c</strong> (1537×2061)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMGTR</td>
<td>1.397</td>
<td>0.319</td>
<td>3.5866</td>
</tr>
<tr>
<td>PMVD</td>
<td>1.489</td>
<td>0.342</td>
<td>3.8452</td>
</tr>
<tr>
<td>SURE</td>
<td>2.104</td>
<td>0.956</td>
<td>11.0</td>
</tr>
<tr>
<td>COLMAP</td>
<td>2.303</td>
<td>1.041</td>
<td>11.704</td>
</tr>
</tbody>
</table>

Moreover, the Herz-Jesu-P8 dataset is used to evaluate the performance of IMGTR on the ground images. The dataset contains eight (8) images of 3072×2048 pixels captured by a Canon D60 digital camera and a ground truth 3D model obtained by laser scanning. Figure 13(a) shows the fifth image of the eight images and Figure 13(b) is the 3D point cloud from laser scanning. Figure 13(c) shows the point cloud generated from the depth of the fifth image in Figure 13(a). It can be observed that the planar surfaces, such as walls can be reconstructed well by IMGTR. However, IMGTR cannot recover curved surfaces well, e.g., at the statues on the wall, due to limited number of triangles in these objects.
The fifth image of the Herz-Jesu-P8 dataset: (a) the image, (b) the ground truth 3D model and (c) the point cloud generated from the depth result of IMGTR.

Figure 13. The fifth image of the Herz-Jesu-P8 dataset: (a) the image, (b) the ground truth 3D model and (c) the point cloud generated from the depth result of IMGTR.

The quantitative evaluation of IMGTR on Herz-Jesu-P8 with reference to several other methods are listed in Table 4. Similar to the evaluation in Hu and Mordohai (2012) and Hou et al. (2018), the percentage of pixels with depth error magnitude less than respectively 2 cm and 10 cm are used while the leftmost (first) and rightmost (last) images in the sequence are excluded. The average depth of ground truth is 12.9 m for a total 36,082,654 pixels. As shown in Figure 13a and Table 4, since there are many curved surfaces in the Herz-Jesu-P8 dataset, IMGTR performs not well without outlier elimination. After outlier elimination, IMGTR(E) shows an accuracy of 77.7% (better than 2cm) and 93.4% (better than 10 cm) for the remaining 20,519,562 pixels, considerably improved compared to LC (Hu and Mordohai, 2012), TYL (Tylecek and Sara, 2010), PMVS (Furukawa and Ponce, 2010), ENZ (Zheng et al., 2014), GIPUMA (Galliani et al., 2015), COLMAP (Schö nberger et al., 2016) and SURE (Rothermel and Wenzel, 2012). However, in comparison to our recent work (Hou et al., 2018), the accuracy is still a few percentage lower even after outlier elimination. On the other hand, as shown in Table 5, comparing to the pixel based dense matching results which are very redundant in the representation of urban scenes, the proposed method is much more efficient, i.e., at most more than 100 times fewer points are needed to represent the scene. Moreover, the accuracy can be improved through reducing the threshold of Harris. However, it will result in inefficiency of scene representation. Thus, the result of IMGTR is a tradeoff between accuracy and representation efficiency. The details of the influence of the threshold of Harris will be discussed in Section 4.3.
Table 4. Quantitative comparison with eight different methods on Herz-Jesu-P8 in terms of the percentage of pixels with error magnitude less than 2 cm and 10 cm, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>&lt;2cm (%)</th>
<th>&lt;10cm (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC (Hu and Mordohai, 2012)</td>
<td>64.9</td>
<td>84.8</td>
</tr>
<tr>
<td>TYL (Tylecek and Sara, 2010)</td>
<td>65.8</td>
<td>85.4</td>
</tr>
<tr>
<td>PMVS (Furukawa and Ponce, 2010)</td>
<td>64.6</td>
<td>83.6</td>
</tr>
<tr>
<td>ENZ (Zheng et al., 2014)</td>
<td>65.0</td>
<td>84.4</td>
</tr>
<tr>
<td>GIPUMA (Galliani et al., 2015)</td>
<td>28.3</td>
<td>45.5</td>
</tr>
<tr>
<td>COLMAP (Schönberger et al., 2016)</td>
<td>69.1</td>
<td>93.1</td>
</tr>
<tr>
<td>SURE (Rothermel and Wenzel, 2012)</td>
<td>58.8</td>
<td>88.1</td>
</tr>
<tr>
<td>PMVD (Hou et al., 2018)</td>
<td>67.6</td>
<td>86.6</td>
</tr>
<tr>
<td>PMVD (E) (Hou et al., 2018)</td>
<td>84.3</td>
<td>97.8</td>
</tr>
<tr>
<td>IMGTR</td>
<td>48.7</td>
<td>66.8</td>
</tr>
<tr>
<td>IMGTR (E)</td>
<td>77.7</td>
<td>93.4</td>
</tr>
</tbody>
</table>

Table 5. Number of points needed to represent a scene with dense points and triangulation

<table>
<thead>
<tr>
<th></th>
<th>Dortmund dataset</th>
<th>Tongchuan dataset</th>
<th>Vaihingen dataset (Area a)</th>
<th>Herz-Jesu-P8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense points</td>
<td>3,013,116</td>
<td>7,140,927</td>
<td>2,531,253</td>
<td>3,406,402</td>
</tr>
<tr>
<td>Triangulation</td>
<td>24,196</td>
<td>28,745</td>
<td>21,091</td>
<td>8,344</td>
</tr>
</tbody>
</table>

4.3 Influence of the Harris threshold and the weights of the energy terms

This section examines the influence of the Harris threshold and the weights of the energy terms with the Herz-Jesu-P8 dataset. As shown in Table 6, while reducing the Harris threshold, the accuracy of the results improves. However, as shown in Figure 14, the image triangulation becomes denser since more corner points are detected, which in turn leads to inefficiency of IMGTR in computational time, data storage and rendering. Thus, very small Harris threshold is not recommended.
Table 6. Influence of the Harris threshold evaluated on Herz-Jesu-P8. The percentage of pixels with error magnitude less than 2 cm and 10 cm, respectively.

<table>
<thead>
<tr>
<th>Harris threshold</th>
<th>Number of Harris points</th>
<th>&lt;2cm (%)</th>
<th>&lt;10cm (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-5}$</td>
<td>5089</td>
<td>77.7</td>
<td>93.4</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>28497</td>
<td>82.3</td>
<td>94.2</td>
</tr>
<tr>
<td>$10^{-7}$</td>
<td>80726</td>
<td>84.5</td>
<td>95.6</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>91771</td>
<td>86.0</td>
<td>96.4</td>
</tr>
</tbody>
</table>

Figure 14. Image triangulations in the red rectangle of the fifth image (a) of the Herz-Jesu-P8 dataset with different Harris thresholds: (b) $10^{-5}$, (c) $10^{-6}$, (d) $10^{-7}$ and (e) $10^{-8}$.

The weights of the energy terms are the smoothness constraint $\sigma$ in Equation 7 and the continuity constraint $\tau$ in Equation 8. As shown in Table 7, when continuity parameter $\tau$ equals 0.1 and 0.5, the accuracy of the results have only minor changes (less than 2%) under a large range of smoothness parameter $\sigma$ (5-30). However, when $\tau$ is changed from 0.5 to 2.5, the accuracy of the results considerably decreases by 5% or more in terms of the errors less than 2 cm, i.e., more pixels tend to have larger depth errors. Thus, we conclude that $\sigma$ is not sensitive while $\tau$ is recommended with a relatively small value. Throughout this study, we use 15 for $\sigma$, and 0.5 of $\tau$. 
Table 7. The percentage of pixels with error magnitude less than 2 cm and 10 cm, respectively under different $\sigma$ and $\tau$ values for the Herz-Jesu-P8 dataset

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$\tau$</th>
<th>$&lt; 2\text{cm} \ (%)$</th>
<th>$&lt; 10\text{cm} \ (%)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.1</td>
<td>77.1</td>
<td>91.7</td>
</tr>
<tr>
<td>15</td>
<td>0.1</td>
<td>77.3</td>
<td>93.3</td>
</tr>
<tr>
<td>30</td>
<td>0.1</td>
<td>76.8</td>
<td>91.2</td>
</tr>
<tr>
<td>5</td>
<td>0.5</td>
<td>77.4</td>
<td>91.8</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
<td>77.7</td>
<td>93.4</td>
</tr>
<tr>
<td>30</td>
<td>0.5</td>
<td>77.0</td>
<td>91.9</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>71.9</td>
<td>91.1</td>
</tr>
<tr>
<td>15</td>
<td>2.5</td>
<td>71.8</td>
<td>92.0</td>
</tr>
<tr>
<td>30</td>
<td>2.5</td>
<td>67.5</td>
<td>91.9</td>
</tr>
</tbody>
</table>

4.4 Limitations

There are some limitations of the proposed IMGTR. As shown in the experiments, IMGTR can reconstruct the flat areas well, but the depth and normal vector estimation of the curved surface is not ideal, since it is hard to describe the curved surface well with limited number of triangles. Another limitation is computational efficiency. It takes 2 hours to reconstruct a dataset with 8 images of 3072 × 2048 pixels in our computation environment (PC with Intel i7-4790 CPU, 3.60Hz and 32GB RAM) without accelerating. The computation of the similarity metric is time consuming and parallelizable with graphical hardware. In addition, even though a relatively low Harris threshold is set while corner points and line segments with LSD (Grompone et al., 2012) are both used to generate image triangulation, it is still difficult to describe the challenging areas well due to the existence of shadows, specular reflectance, occlusion, etc. Proper image pre-processing, e.g. image enhancement could possibly be helpful to better reconstruct these areas.

5. Conclusion

This paper introduces an image-triangle based multi-view 3D reconstruction (IMGTR) method for urban scenes. With the input of multi-view images and corresponding camera parameters, IMGTR can produce the depth map and normal vector map for each image. Starting from constructing triangulation with line segments and corner points for each images, IMGTR reconstructs the 3D information of the
scene through finding the optimal object planes of the triangles in each image under the image similarity constraint, the smoothness constraint and the continuity constraint on the edges between adjacent triangles. To minimize the energy function, graph cut technique and an optimization framework similar to the EM method are used. After the optimization, the depth maps and normal maps of every image can be obtained.

IMGTR performs well on urban scenes due to several formulations and optimization strategies. Due to the planar geometric model, the similarity metric is computed according to the image triangle instead of the fronto-parallel support window. The determination and carefully handling of the depth-fracture avoid the noises or over-smooth in the discontinuous areas. Besides, the utilization of the extended PatchMatch (Hou et al., 2018) enables IMGTR to achieve acceptable results even with random initialization. Furthermore, the representation of the urban scenes is more efficient compared to pixel based dense matching method since a density-adaptive image triangulation can be constructed with line segments and corner points.

Four datasets are used to evaluate the proposed IMGTR. Qualitative experiment shows that IMGTR can reconstruct the scene geometry with higher fidelity, especially in the discontinuous edge areas, than COLMAP and PMVD. In addition, IMGTR achieves an expected density adaptive reconstruction, i.e., in areas with rich texture and complex structure, details are described with dense triangles, while in areas with simple structure, the surfaces are expressed with larger triangles. Quantitative experiments with the Vaihingen dataset show that IMGTR achieves the highest accuracy compared with SURE, COLMAP and PMVD. Specifically, IMGTR shows an average improvement of 4.27% in position accuracy compared to PMVD, 50.94% compared with SURE and 54.76% compared to COLMAP. More beneficially, the reconstructed scenes have higher fidelity in planar urban structures and can achieve density adaptive reconstruction, which means the triangles are dense in complex areas while they are sparse in planer areas. It should be noted that the reconstruction of IMGTR is robust to initial noisy normal map and is suitable for urban scenes. The experiments with the Herz-Jesu-p8 dataset show that IMGTR cannot reconstruct the curved surfaces well. However, IMGTR can represent the reconstructed scenes more efficiently, i.e. more than 100 times fewer points are needed to represent a scene compared with dense point clouds. Thus, it is highly recommended to use IMGTR to reconstruct urban scenes of ground and planar building roofs.

Since the images used for reconstruction are of high resolution, e.g. nadir ground sampling distance of 6
cm for the UAV images, ground sampling distance of 8cm for the Vaihingen dataset, the effectiveness of airborne images with a resolution of a few centimeters has been demonstrated. However, the generalization to satellite images might need further investigation. Several factors should be considered. Despite the resolution of satellite images is getting close to the ones from air, their contrast is often not ideal considering the sensor sensitivity, atmospheric conditions, the quality of radiometric calibration, etc. Moreover, small base-height (B/H) ratio (comparing to aero images) or intersection angles of satellite images would lead to high geometric uncertainty for depth map generation. Finally, it is uncommon that we have more than a few overlapping satellite images over the same area. Investigating these issues can be interesting future work.

References


and Pattern Recognition. https://doi.org/10.1109/CVPR.2010.5539804


