Integrated Image Matching and Segmentation for 3D Surface Reconstruction in Urban Areas

Lei Ye and Bo Wu

Abstract
High-resolution imagery, which features the advantages of high-quality imaging, a short revisit time, and lower costs, is an attractive option for 3D reconstruction applications. Photogrammetric 3D reconstruction requires reliable and dense image matching. In urban areas, however, image matching is particularly difficult because of the complexity of urban textures and the severe occlusion problems caused by buildings. This paper presents an integrated image matching and segmentation approach (named SATM+) for 3D reconstruction in urban areas. SATM+ is based on our existing self-adaptive triangulation-constrained matching (SATM) framework and incorporates three novel aspects to address image matching challenges in urban areas: (1) image segmentation-based occlusion filtering, (2) segment-adaptive similarity measurement to reduce matching ambiguity, and (3) local and regional dense matching propagation to generate reliable and dense matches. We performed an experimental analysis of two sets of high-resolution urban images, and the 3D point clouds generated using the proposed SATM+ were compared with airborne light detection and ranging (lidar) data and the point clouds generated using the semi-global matching (SGM) method. The results indicate that SATM+ can generate 3D point clouds with a geometric accuracy comparable to that of lidar data but a much higher point density. SATM+ performs similarly to SGM in relatively flat areas, but is superior in built-up areas. The proposed approach is a promising option for image-based 3D surface reconstruction in urban areas.

Introduction
In recent decades, urban 3D modelling has emerged as an important issue in the fields of photogrammetry and computer vision, with various applications in urban planning, urban monitoring, and urban management. Although urban 3D data are in high demand worldwide, these data are still generated using rudimentary and relatively expensive methods (Gruen, 2008; Vanegas et al., 2010; Zhu et al., 2010). To date, two main techniques have been used to generate 3D data: airborne light detection and ranging (lidar) techniques and image-based photogrammetric techniques.

Since the 1990s, lidar techniques have been increasingly used to collect 3D urban data (Gamba and Houssmand, 2000). The current systems allow the measurement of approximately four surface points per square meter, with a vertical accuracy of approximately 15 cm (Hyyppä et al., 2000). It should be noted, however, that airborne lidar surveys are usually expensive (Shan and Toth, 2008).

Image-based photogrammetric techniques have been widely used for 3D reconstruction due to the advantages of high-quality imaging, a short revisit time, and lower costs. A successful photogrammetric 3D data derivation requires automated, reliable, and dense image matching. The techniques used for image-based 3D reconstruction have advanced since the late 1980s (Ackermann and Krzystek, 1991), and suitable image matching algorithms can be used to generate 3D point clouds with favorable levels of accuracy, reliability, and detail in relatively favorable texture conditions (Gruen, 2008; Bleyer et al., 2011; Wu et al., 2011 and 2012). However, image matching in urban areas is particularly difficult. Most traditional digital photogrammetry systems require considerable human labor to process images in urban areas (Helpke, 1995; Wu et al., 2011), especially in metropolitan regions containing densely packed areas of vast skyscrapers and tall buildings. These challenges are mainly attributed to the intrinsic problem of image matching, which is caused by the complexities of urban textures and severe occlusion problems associated with buildings.

However, the textural complexity of urban areas can actually increase the accuracy and convenience of feature-based image matching, assuming an effective matching strategy is available. The high density of buildings tends to offer regularly shaped image segments. These advantages could potentially facilitate image-based 3D surface reconstruction in urban areas.

The present study used high-resolution imagery of urban areas for 3D reconstruction, with the aim of developing a cheaper, better automated image-based 3D reconstruction method. Following the literature review in the next section, we present our novel integrated image matching and segmentation approach to 3D surface reconstruction in urban areas. Then, we describe the results of experiments performed using two sets of high-resolution images, each representing a typical urban type. Finally, the concluding remarks are presented and discussed.

Related Works
Image matching is used to identify corresponding pixels between images that can be used for 3D reconstruction using photogrammetric space intersection (Wu et al., 2012). Image matching is also a difficult, essential task in the fields of photogrammetry and computer vision (Lhuillier and Quan, 2002; Zhang and Gruen, 2006; Wu et al., 2011). Currently, two image matching strategies are normally used: sparse point matching, which is often used for image registration, and dense matching, which is often used for 3D reconstruction.

Sparse point matching normally includes interest point detection and matching. The interest point detectors widely used in photogrammetry include the Moravec detector (Moravec, 1981), the Förstner detector (Förstner, 1986), and the Harris detector (Harris and Stephens, 1988). The Moravec detector measures the gray value differences between a window and windows shifted in several directions, and detects interest points if the minimum of these differences is superior.
to a threshold (Moravec, 1981). The Förstner detector extracts interest points by searching for optimal windows using the auto-correlated matrix and optimizes the point locations based on a differential edge intersection approach ( Förstner, 1986). The Harris detector improves the Moravec detector by using an auto-related matrix to ensure better repeatability (Harris and Stephens, 1988). Traditional software packages equipped with these detectors generally used the normalized correlation coefficient as the matching metrics. Various constraints have been employed to improve the matching performance, such as the epipolar constraint (Gruen and Baltsiavas, 1998), triangulation constraint (Wu et al., 2012), and redundant information from multiple images (Agouris and Schenk, 1992; Zhu et al., 2010). The point matches can be optimized to sub-pixel level by using the least-squares matching methods (Gruen, 1985 and 2012).

Since the milestone work of the scale invariant feature transform algorithm (SIFT) (Lowe, 1999), the use of invariant features has offered another paradigm for sparse point matching. SIFT combines a scale invariant interest point detector and a descriptor based on the gradient distribution in the detected local regions (Lowe, 2004). SURF, which was inspired by the SIFT algorithm (Bay et al., 2008), incorporated the Hessian matrix approximation to speed up matching. Other algorithms that use machine learning techniques to detect features, such as SUSAN (Smith and Brady, 1997) and FAST (Rosten and Drummond, 2006), have also been used widely for point matching. Despite the subsequent introduction of many new developments, the SIFT has proven remarkably successful, and most subsequent improvements to it were designed for efficiency.

In the past, a large number of dense matching methods have been developed to suit different contexts. For dense matching, disparity computation is the most important step. Methods used to calculate disparity can be classified into three main categories: local methods, global methods, and methods based on a combination of local and global factors. Local methods simply select the disparity with the minimum cost value; i.e., a local “winner-take-all” optimization. These local methods feature different strategies, such as block matching and gradient-based optimization. Block matching methods calculate disparity by seeking the highest correlation between the reference image and the matching image within a region around the pixel of interest. The most significant drawback of block matching methods is sensitivity to locally ambiguous regions in images. Gradient-based methods determine small local disparities between two images by formulating a differential equation that correlates motion and image brightness. Notably, local methods share a common limitation: the uniqueness of matches is only enforced for the reference image, whereas points in the searching image might be matched to multiple points (Stentoumis et al., 2014).

In contrast to local methods, global methods have been formulated in an energy-minimization framework (Geiger et al., 2010; Shaobo et al., 2012; Blumenthal-Barby and Eisert, 2014). These methods aim to identify a disparity function that minimizes global energy. The global methods incorporate different strategies, including dynamic programming, belief propagation, and graph cut. Dynamic programming approaches (Van Meerbergen et al., 2002; Veksler, 2005) involve the calculation of a minimum-cost path through the matrix of all pairwise matching costs between two corresponding scanlines. Belief propagation approaches (Sun et al., 2003; Yang et al., 2009) attempt to solve the maximum posteriori Markov Random Field problem using global optimization. The graph cut method (Boykov et al., 2001) was proposed based on the max flow algorithm in graph theory. Local methods are considered more straightforward and efficient for real-time applications and large dataset manipulation. However, these methods are inferior to others in terms of accuracy and robustness. The elaborate models used by global methods to describe the matching process tend to increase computational loads. Therefore, combinations of local and global methods, which retain the strengths while discarding the weaknesses of each technique, are now considered the most successful option for image matching. Semi-global matching (SGM) (Hirschmuller, 2011) is the most widely used of these combination methods.

SGM was initially proposed to handle complex image matching and uses mutual information (MI) as the cost function (Viola and Wells III, 1997). MI depends on the individual and joint entropies of two images. SGM uses the global cost function with a smoothness constraint that penalizes discontinuities. In contrast to the original version, the latest cost function separates the penalty function into two parts: one with pixel discontinuities, and one with larger discontinuities (Hirschmuller, 2011), as shown in Equation 1.

\[
E(D) = \sum_{p} C(p, D_p) + \sum_{q \in N_p} R_q^T [D_p - D_q] + 1 \cdot \sum_{q \in N_p} R_q^T [D_p - D_q] > 1
\]  

This cost function performs well in flat or radiometrically uniform areas, or even in areas with lower building densities. However, the cost terms \(C(p, D_p)\) may decrease while the penalty terms \(\sum_{q \in N_p} R_q^T [D_p - D_q] > 1\) increase in areas with high densities of buildings. As a result, the matching performance in these areas may not be as good as expected (with reference to the experimental analysis in this paper). Regarding the image matching methods described above, matching difficulties in urban areas, which are mainly caused by the occlusion problems from buildings, have rarely been considered. In most image matching methods, areas of occlusion are addressed by detection after matching. Some research studies have compensated for occlusion problems by using hard constraints, such as the uniqueness constraint (Kolmogorov and Zabih, 2001) and ordering constraint (Silva and Santos-Victor, 2000). The uniqueness constraint enforces one-to-one mapping between pixels in two images, while the ordering constraint expresses two or more along scanlines. However, the abilities of these constraints to solve occlusion issues are limited. The computer vision community has also used image segmentation to estimate disparities in close-range images (Bleyer and Gelauxt, 2005; Bleyer et al., 2011). The resulting disparity maps are normally interpolated from the matching results and are relatively smooth. However, images in urban areas are very complicated. Disparities vary drastically between buildings, and may not be smooth even within a single building. Therefore, a direct disparity interpolation is not rigorous.

This paper proposes an integrated image matching and segmentation approach. By incorporating image matching with image segmentation, objects on images with significantly different disparities will be handled separately. The image segmentation results will be used for occlusion filtering and similarity measurements to provide better matching results for 3D surface reconstruction in urban areas.

**Integrated Image Matching and Segmentation Approach**

The integrated image matching and segmentation approach, named SATM+, is based on our existing self-adaptive triangulation-constrained matching (SATM) framework (Wu et al., 2011, 2012). In the SATM framework, the first step is to identify several robust seed points on stereo images, which are used to...
construct a pair of initial triangulations on the images. Next, feature points are matched based on the triangulation constraint and other constraints. The newly matched points are inserted into the triangulations, so that they are densified dynamically along with the matching propagation. Because the points with the most distinct textures are always the easiest and earliest to be matched, the dynamic updating of triangulations is self-adaptive to image textures. Accordingly, it propagates the constraints from robust matches with favorable textures to the matching in areas with less favorable textures using triangular segmentation. Finally, reliable matching results can be obtained.

Wu et al. (2012) further incorporated edge matching into the SATM framework. However, the application of SATM to images of urban areas causes severe problems due to the aforementioned matching difficulties. Therefore, the SATM+ proposed in this paper incorporates the following new components in an attempt to address the challenges of image matching in urban areas: (1) occlusion filtering based on image segmentation, (2) segment-adaptive similarity measurement to reduce matching ambiguity, and (3) local and regional dense matching propagation to generate dense and reliable matches.

Overview of the Approach
Figure 1 illustrates the workflow of the SATM+ approach. Here, robust feature matches for a pair of high-resolution images of urban areas are obtained using the previous SATM approach (Wu et al., 2011 and 2012). Image segmentation of one of the stereo images is then performed. The feature matches are used to construct a pair of corresponding triangulations on the images, and the disparities of the matched feature points are used to interpolate a disparity map. The image segmentation results are used to fit the disparity map to planes with constant disparities (e.g., flat building roofs or ground patches) or smoothly varying disparities (e.g., building facades). Meanwhile, image matching is conducted under triangulation constraint and other constraints (e.g., epipolar constraint). Once a matching candidate is obtained, occlusion filtering is performed to determine whether the disparity fits the surrounding disparity plane. If the fit is good, the matching candidate is accepted as a successful new match and inserted immediately into the triangulations. The surrounding disparity plane is also updated simultaneously by adding the disparity of the new match, and the corresponding triangulations and disparity planes are dynamically updated along with the matching propagation. During the matching process, a segment-adaptive similarity measurement is used to reduce matching ambiguity in areas near segment boundaries.

The completion of the above-described integrated image matching and segmentation process yields a pair of densified corresponding triangulations from the final matched results, as well as updated and more accurate disparity planes. These outputs are subsequently used to provide constraints for the next dense matching step. Using the dense matching results, 3D point clouds can be derived using photometric space intersection based on the image orientation parameters, and the 3D surface reconstruction results (e.g., digital surface models (DSMs)) can be generated.

Image Segmentation Based on Edge Detection and Region Growing
Image segmentation is a primary component of the proposed approach. In high-resolution images of urban areas, buildings are represented by roofs and facades and normally have regular shapes with sharp edges. Our approach first uses a well-known edge detection algorithm to detect the edges in images. Subsequently, closed edges are used to form segments, and a region growing algorithm is used to reduce the over-segmentation problem.

For edge detection, we used the EDSION (edge detection with embedded confidence) algorithm (Meer and Georgescu, 2001), which is more stable and accurate than other algorithms used for high-resolution remote sensing images (Li et al., 2010). The EDSION algorithm includes three steps (gradient
estimation, non-maximum suppression, and hysteresis thresholding) and can detect weak edges without introducing spurious edges (Li et al., 2010). For edges detected using the EDSION, a local re-weight strategy (Arbelaez et al., 2009) is used to detect the local maximum near the endpoints of the edges until the edge intersects with a neighborhood edge. Finally, the closed edges are obtained to form segments.

However, our extensive experiments with urban images suggest that the exclusive use of the above method leads to significant over-segmentation problems due to the complex textures in urban images. Therefore, a region growing algorithm (Felzenszwalb and Huttenlocher, 2004) is used to reduce over-segmentation by merging and growing regions with similar properties.

Notably, image segmentation is not the emphasis of this research, and therefore the well-known segmentation algorithms were used. It might be necessary to interactively edit the automatically obtained segmentation results to ensure reliability and effectiveness. Figure 2 presents an example of the segmentation results.

Occlusion Filtering
In stereo vision, the occlusion problem refers to the visibility of some parts of the scene by one camera but not the other camera as a consequence of scene and camera geometries. Urban areas contain many skyscrapers and tall buildings that present opportunities for occlusion (e.g., object self-occlusion and occlusion between multiple objects) (Xing et al., 2009). To reduce the ambiguities caused by occlusion problems, we propose an occlusion filter in the image matching framework based on previous image segmentation results.

Occlusion filtering assumes that disparities will change smoothly within the same plane (either a building roof or facade), but will change suddenly when the image point changes from a facade location to a roof and vice versa. Therefore, two scenarios are considered specifically: (1) candidate match within a segment, and (2) candidate match outside a segment but within a buffer zone of the segment boundary. Figure 3 shows a flowchart of the occlusion filtering method. An initial disparity map is generated based on feature matching results obtained during the previous step. The image segmentation results and RANSAC algorithm (Fischler and Bolles, 1981) are used to fit the disparities \( D \) and the image coordinates \((u,v)\) within each segment, using a robust plane model (e.g., for the building roof) or a quadratic surface model \( D = S(u,v) \) (e.g., for building facades). From the established plane model, each pixel \((u_v,v)\) within the segment will be able to have an estimated disparity \( D_{est} = S(u,v) \) based on interpolation. For the first scenario of candidate match within the segment, if the disparity from the candidate match \( D_{C} = S(u',v') \) is within a threshold range of the interpolated disparity \( D_{est} \), the match is considered successful. Disparities beyond the threshold range suggest a mismatch possibly due to occlusion problems that should be excluded. For the second scenario of candidate match outside a segment but within a buffer zone of the segment boundary, the disparity of the candidate match \( D_{C} \) should be outside the threshold range, and a failure to meet this criterion may also indicate a mismatch. The newly matched point \((u,v,D_{C})\) will be collocated, and used to re-fit the plane dynamically by RANSAC, and inserted into the triangulations dynamically to ensure that the plane will gradually approach the real situation. Notably, occlusion filtering sets a loose constraint on earlier stages of matching propagation when the smaller number of matched points reduces the accuracies of the disparity planes. These constraints tighten as the number of matched points increases during the later stages of the matching propagation.
In contrast to other post-filtering strategies, this occlusion filtering process is incorporated into the matching propagation. Particularly, this occlusion filtering method not only helps to remove mismatches from occluded areas during matching, but also provides clues for corrected matching, thus leading to improved matching in occlusion areas.

Figure 4 shows an example of matching results obtained with/without the occlusion filtering. Some points around the buffer zones of the building roofs were matched incorrectly because of their similar intensities. These mismatches were eliminated by applying the occlusion filtering, and the matched results coincided well with the building roof boundaries.

**Segment-Adaptive Similarity Measurement**

In image matching, the matching cost is normally exploited to compare the relationships or similarities of the corresponding pixels and thus determine a correlation coefficient. However, the correlation assumes equal depths for all pixels within a local window. Notably, this assumption violates situations with occlusion problems caused by depth discontinuities. Figure 5 illustrates two types of typical decision conflicts regarding similarity measurement that occur during the search for matches around building boundaries. In this case, the correlation window must be split into two parts. One type of decision conflict is a “border defect”, wherein parts $a$ and $b$ correspond to the two correlations $CC_a$ and $CC_b$. The concept that part $a$ owns the higher correlation coefficient is easily confirmed. However, the similarity measure of part $b$ is lower due to a mis-correspondence. As a result, the overall correlation may not exceed the predefined threshold. This may lead to missing of correct matches near the borders of the area. The other type of decision conflict is “border blurring”. In Figure 5, the right-hand image (after splitting the correlation window into two parts) contains three parts. However, parts $b_1$ and $b_2$ exhibit a mis-correspondence, leading to a higher similarity for part $a$ than for parts $b_1$ and $b_2$. As a result, the overall correlation may exceed the predefined threshold, causing extension of the roof beyond the building and blurred border.

Inspired by the segmentation-based occlusion filtering, a segment-adaptive similarity measurement method is introduced. Figure 6 illustrates the basic concept underlying the method. Assuming the grey region in Figure 6a represents the...
foreground (e.g., building roof) and the white region represents the background (e.g., the ground), the line between the gray and white regions marks the border between the foreground and background. The box with dashed lines indicates the matching window. According to the segmentation and correlation principle, the correlation calculation should involve the

Figure 5. Two types of decision conflicts regarding similarity measurement.

Figure 6. Segment-adaptive similarity measurement for point matching: (a) A matching window separated by an edge, and (b) weight distribution in the matching window.
grey pixels, whereas the white pixels may cause a decrease in the correlation or even a mismatch. In principle, the white area should be invalidated when calculating the correlation. In practice, however, a buffer zone along the edge in the matching window should be considered because of the limitations in image resolution and edge detection accuracy, and pixels within the buffer zone should also be involved in the correlation calculation. In accordance with the above observations, the segment-adaptive similarity measurement uses a sigmoid function to assign weights to each pixel in the matching window when calculating the correlation coefficient. These weights depend on the distance from the pixel to the edge.

The sigmoid function is applied to weight the pixels in the matching window as follows:

$$S(t) = \frac{1}{1 + e^{-\alpha t}}$$

where $t$ represents the signed distance from the pixel to the edge. If the pixel is on the same side as the point to be matched, $t$ is positive; otherwise, it is negative. $\alpha$ is the coefficient used to determine the size of the buffer zone, based on the image resolution and the edge detection accuracy. Figure 6b shows an example of the weight distribution for pixels within the matching window shown in Figure 6a. The black line in the middle of Figure 6b corresponds to the edge between the white and grey regions in Figure 6a.

Figure 7a and 7b present a subset of paired high-resolution satellite images taken near a building. The black cross in the left image is the point to be matched in the right image. The segment-adaptive similarity correlation values within the search range (disparity: ±25 pixels near the predicted disparity) are shown in Figure 7c, and the traditional normalized correlation coefficient (NCC) values are shown in Figure 7d. According to the matching strategy, the pixels with the maximum correlation coefficient values are candidate matches, and are labeled with black (segment-adaptive similarity correlation) and white crosses (traditional NCC) in the right image. As depicted in Figure 7a and 7b, the black cross is the correct match. Figure 7c and 7d depict many negative correlation coefficients in the traditional NCC, whereas all correlation coefficients for the segment-adaptive similarity are positive. We attribute this to the fact that in the left image, the facade near the feature point is brighter than the roof, whereas the facade is occluded in the right image. The traditional NCC still includes the pixels from the facade in the matching window, whereas the corresponding pixels in the right image are located on the ground and are darker than the roof pixels, leading to negative
correlation coefficients in this region. However, when using the segment-adaptive similarity correlation, the pixels within the roof region receive significantly greater weights in the similarity measurement. Accordingly, the pixels on the ground have a weaker influence on the similarity measurement.

**Dense Matching Propagation**

Based on the feature matching results from the previous steps, a dense matching propagation is carried out to obtain dense matching results for subsequent 3D surface reconstruction. The dense matching propagation includes the following two strategies: local matching propagation and regional matching propagation. Figure 8 provides a framework for the dense matching propagation.

First, a seed match list comprising all vertices of the triangulations from the previous feature matching is generated. The points in the seed list are sorted according to their correlation coefficients in feature matching. The dense matching propagation begins with the most reliable match in the seed list and propagates the matching of this point to its neighborhood using the local matching propagation strategy. The newly matched points are then inserted into the seed list based on their correlation coefficients. Subsequent matches in the seed list are then selected for matching propagation until all points have been processed. Some regions in the images may remain unmatched (i.e., cavities) after local matching propagation and are subjected to regional matching propagation. The aforementioned disparity planes, segment constraint, and segment-adaptive similarity measurement are used for both local and regional matching propagation. The final output is a set of dense matching results from which photogrammetric point clouds and DSMs can be generated.

The local matching propagation follows the “best-first” strategy used in popular dense matching methods (Lhuillier and Quan, 2002; Megyesi and Chetverikov, 2004; Zhang and Gruen, 2006). In this strategy, the most reliable seeds are used to direct the matching propagation process. For a listed seed located in a pre-defined segment, the disparities of the 24 connected neighboring pixels (a $5 \times 5$ neighborhood) $u$, in one image will be interpolated from the corresponding disparity plane. If the seed is located beyond any pre-defined segment, the disparity of the seed itself will be assigned to its neighboring pixels. The match candidates of the neighboring pixels in the other image $u'$ can then be estimated based on the disparities. For each neighboring pixel $u_i$, a search range $(u_i - d_i, u_i + d_i)$ centered at $u_i$ along the epipolar line will be examined; here, $d_i$ is a pre-defined minimum searching distance. The pixel with the highest correlation coefficient within the search range is labeled as a match. Once a newly matched pair of points is obtained, it is inserted into the seed list according to the correlation coefficient. If all 24 neighboring pixels of a seed have been processed, the seed will be removed from the list. After all of the listed seeds have been processed, the local matching propagation is terminated.

Notably, the local matching propagation incorporates the segment constraint. In other words, if some of the 24 neighboring pixels of a seed are beyond the segment, matching propagation will stop at those pixels. The previously described segment-adaptive similarity measurement is also used to determine the correlation coefficients.

Local matching propagation has the following advantages. First, its use guarantees robust and stable matching, as the robust matches are propagated before others according to the “best-first” strategy. Second, the disparity constraint facilitates the matching of points in highly textured areas to those in poorly textured areas. However, local matching propagation occurs in immediate neighboring regions and may be cut off by the coefficient threshold, possibly leading to the formation
of cavities corresponding to some unmatched pixels, as shown in Figure 9a. Therefore, a regional matching propagation is designed to facilitate dense matching in the cavities.

For the regional matching propagation, the previous segmentation result is used to check whether an unmatched pixel in a cavity in one image is located in any segment. If the pixel is located inside a segment, the interpolated disparity from the disparity plane is used to predict a candidate match in the other image, and the matches will be searched within a range along the epipolar line, as previously described. If the pixel is not located inside any segment, its disparity will be estimated based on a clustering analysis of the disparities of existing matches within a buffer region. In this case, the disparity estimation is implemented as follows:

1. For each unmatched pixel \( P \), the existing matches are searched within a buffer surrounding region, and their disparities are classified using the k-means algorithm (MacQueen, 1967);
2. The statistical feature of each disparity cluster is analyzed. For each cluster, the mean value \( \overline{d}_j \) and the standard deviation \( s_j \) are calculated (\( j \) represents the \( j \)th cluster);
3. To match pixel \( P \), the disparity range \( [\overline{d}_j - 2s_j, \overline{d}_j + 2s_j] \) is searched to find the most possible match for each cluster \( P' \) with the maximum correlation coefficient \( CC_j \);
4. The status of the correlation coefficient with regard to the threshold is checked. If no \( CC_j \) passes the threshold, the pixel \( P \) is not matched. If only one \( CC_j \) passes the threshold, pixel \( P \) is matched with \( P' \). If more than one \( CC_j \) passes the threshold, the pixel with the highest \( CC_j \) is selected.

Following local and regional dense matching propagation, all pixels in an image have been checked for possible matches and all of the regions that could not be matched have been isolated. Figure 9 shows an example of the regional matching propagation results for the same area of Figure 9a. Here, the regional matching propagation effectively matched the pixels in the cavities. Although a few regions in Figure 9b remain without matches, these are mainly attributable to issues such as occlusion, moving objects, or insufficient textural information.

**Experimental Analysis**

To evaluate the proposed approach, we conducted an experimental analysis of two sets of high-resolution images, each representing a different urban type. The first set comprised aerial images of Vaihingen, Germany that were selected from ISPRS benchmarks, and the second set is a pair of Pleiades-1 satellite images of Hong Kong.

After image matching, 3D point clouds can be obtained using a photogrammetric space intersection based on the pin-hole camera model (for aerial images) or the rational polynomial coefficient (RPC) model (for satellite images). The generated 3D point clouds were evaluated in terms of their geometric accuracy and other aspects (e.g., point density and distribution) and subsequently compared with 3D point clouds from the airborne lidar data. During the quantitative evaluation, the 3D coordinates of the generated point clouds were shifted to the reference airborne lidar point cloud to ensure that the evaluation was not biased by possible errors in the image orientation parameters and the lidar point cloud itself. To compare the photogrammetric and lidar point clouds, a 3D triangulated mesh model was generated first from the photogrammetric point cloud, and a new set of heights of the lidar points were interpolated from the 3D mesh model. They were then compared with the corresponding heights of the lidar points, and their signed differences were computed. The
root-mean-square-error (RMSE) and standard deviation (STD) values of the signed differences were calculated. Similarly, 3D point clouds were obtained from the matching results generated by SGM and were also compared with those from the proposed approach.

**Experiment Based on Aerial Images of Vaihingen, Germany**

The first data set of images captured over Vaihingen, Germany was obtained from the ISPRS 3D Building Reconstruction Benchmark. The images had a ground sampling distance of 8 cm (above-ground flying height: 800 m, focal length: 120 mm, 65 percent forward lap, 60 percent side lap) and were taken with an Intergraph/ZI DMC camera. To investigate the performance of the approaches in urban areas, a built-up region (the rectangle in Figure 10) was selected for the experimental analysis.

The reference dataset used in this experiment was a 3D point cloud generated from airborne laser scanning data. They were acquired using a Leica ALS50 system with a 45° field-of-view and a mean above-ground flying height of 500 m. The average strip overlap is 30 percent, and the point density varies between 4 and 7 pts/m².

Figure 11a shows part of the point clouds generated using the proposed SATM+ method (colored small dots) and lidar (white square dots). For some regions (e.g., near the buildings), SATM+ could not produce 3D points, whereas the lidar points...
are evident. As these regions are occluded, no matches can be generated by image matching. In other areas, SATM+ could generate more dense 3D point clouds relative to the lidar data. Figure 11b is a higher-magnification view of the box in Figure 11a. The lidar points (in white) are enlarged seven-fold for better visualization. Here, the point density obtained by SATM+ is clearly much higher than the lidar data. The statistics indicate that on average, the density of the point clouds obtained by SATM+ is approximately 13 times higher than that of the lidar point clouds.

Figure 12 depicts the shaded DSMs interpolated from 3D point clouds derived using the proposed SATM+, SGM, and lidar data. The SGM appeared to yield a slightly smoother DSM, compared to that generated by SATM+. However, the latter offered sharper building boundaries (e.g., the vertical walls of buildings in the enlarged views in Figure 12 compared with the former. This can be attributed directly to the integrated image matching and segmentation strategy, which yields improved matching performances in built-up areas.

The results of the statistical evaluation are shown in Table 1. According to the features in this Table, the RMSE and STD values for SGM were slightly higher than those for the SATM+. However, the two image matching methods yielded similar point densities. Notably, both methods provided greater densities than the lidar point clouds.

Table 1. Statistical analysis of the Vaihingen Experiment.

<table>
<thead>
<tr>
<th></th>
<th>RMSE (m)</th>
<th>STD (m)</th>
<th>Density (pts/m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGM</td>
<td>2.104</td>
<td>2.102</td>
<td>101</td>
</tr>
<tr>
<td>SATM+</td>
<td>1.572</td>
<td>1.567</td>
<td>106</td>
</tr>
</tbody>
</table>

Experiment Based on Pleiades-1 Satellite Images of Hong Kong

Hong Kong is renowned for its impressive skyline, which comprises a very high density of skyscrapers. In this study, we used Pleiades-1 satellite images to evaluate the performance of matching methods in metropolitan areas, such as the Central District of Hong Kong. A pair of Pleiades-1 stereo images was acquired on 04 March 2013. These images have a ground resolution of 0.5 m/pixel, and the pair has a convergence angle of 14.8°. A typical densely built-up area was selected for study (the rectangle marked in Figure 13).

The airborne lidar data used for reference was collected between 01 December 2010 and 08 January 2011, and covered the Central District of Hong Kong. The vertical accuracy of the lidar data is about 10 cm and the horizontal accuracy is about 1 m according to the metafile. The point density is about 4 pts/m².

Figure 14 shows shaded DSMs of the study area that were generated using different approaches. In the flat region in the northern part of the study area, the DSM yielded by SGM was smoother than that yielded by SATM+. By contrast, in the south and east parts comprised densely packed buildings with severe occlusions, only the SATM+ recovered most of the tall buildings. We attribute this outcome to differences in the methods; in SATM+ the matching propagation begins with feature extraction, whereas SGM uses global optimization, which has been found to contribute to over-smoothing.
The results of a statistical evaluation are shown in Table 2. In this experiment, all statistical values indicated the superiority of SATM+ relative to SGM. However, the relatively lower resolution (0.5 m/pixel) of the satellite images yielded point densities that were inferior to the lidar data.

Table 2. Statistical analysis of the experiment in Central District, Hong Kong.

<table>
<thead>
<tr>
<th></th>
<th>RMSE (m)</th>
<th>STD (m)</th>
<th>Density (pts/m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGM</td>
<td>6.809</td>
<td>6.015</td>
<td>2.1</td>
</tr>
<tr>
<td>SATM+</td>
<td>5.509</td>
<td>5.417</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Visually, the DSM generated by SATM+ may not be as appealing as the lidar DSM, as shown in Figure 14. Notably, however, the DSM generated by SATM+ only used two images, and only the objects visible in both the two images could be reconstructed. If multiple image views are available, the proposed method could be used to generate 3D surface models with full structures.

Conclusions and Discussion

The approaches presented in this paper address 3D surface reconstruction in urban areas, the most complicated challenge in the field of photogrammetry. The theoretical analysis and experimental validation presented in this paper demonstrate the performance of the developed approaches and yield the following conclusions:

1. The integrated image matching and segmentation approach helps to alleviate matching ambiguities in urban images, which are caused by the existence of buildings. Here, occlusion filtering can eliminate the mismatches caused by occlusions and improve the matching reliability. The segment-adaptive similarity measurement can compensate for defects caused by discontinuous disparities.

2. The proposed dense matching strategies, including local and regional propagation, are used to propagate matching from the robust feature matches to other regions of the images. These dense matching strategies enable an evaluation of all pixels for possible matches and the isolation of regions without any matches.

3. According to the experimental analysis, the proposed SATM+ performs better than a popular matching method (i.e., SGM) in terms of the preservation of feature boundaries and recovery of tall buildings. These advantages are largely attributable to feature-based matching and the integrated matching and segmentation. Our results demonstrate that SATM+ is able to generate 3D point clouds comparable to or superior to those by lidar.

The research and developments presented in this paper offer an alternative option for 3D surface reconstruction in urban areas. Our developed approaches appear to be significant for urban 3D reconstruction and modelling applications. It should be noted that, this paper focuses on stereo image matching for 3D surface reconstruction. Using multiple images (three or more) has the potential to further improve the matching performances by taking advantage of the redundant information, which will be our future efforts. Future works will also include improvements to image segmentation, optimization of the algorithms, and reductions in processing time.

Acknowledgments

This work was supported by grants from the Research Grants Council of Hong Kong (Project No. PolyU 5330/12E and Project No. PolyU 152086/15E), and a grant of strategic importance from the Hong Kong Polytechnic University (Project No. 1-ZE24).
Figure 14. Shaded DSMs of Hong Kong, generated using different approaches.

References
Bleyer, M., C. Rheinmuth, and C. Rother, 2011. PatchMatch stereo- 

to-neutral image-based modeling, Computers & Graphics 

39:89–100.


minimization via graph cuts. IEEE Transactions on Pattern 

Analysis and Machine Intelligence, 23:1222–1239.


image segmentation. International Journal of Computer Vision, 


paradigm for model fitting with applications to image analysis 

and automated cartography. Communications of the ACM, 

24:381–395.

 Förstner, W., 1986. A feature based correspondence algorithm for 

image matching. International Archives of Photogrammetry, 


Gamba, P., and B. Houssinam, 2000. Digital surface models and 

building extraction: A comparison of IF SAR and LIDAR data. 


1968.


matching. Proceedings of the Asian Conference on Computer 


matching technique. South African Journal of Photogrammetry, 


multiphoto matching. Photogrammetric Engineering & Remote 


Harris, C., and M. Stephens, 1988. A combined corner and 

dependency detector. Proceedings of the Alvey Vision Conference, 

Manchester, UK, p. 50.


workstations for topographic applications. Photogrammetric 


Hirschmuller, H., 2011. Semi-global matching-motivation, 

developments and applications, Proceedings of Photogrammetric 

Week. Hyyppae, J.M., U. Pyysalo, H. Hyyppae, H.G. Haggren, and 

generation in forested areas, Proceedings of Proceedings of 

International Society for Optics and Photonics, pp. 119–130.


correspondence with occlusions using graph cuts, Proceedings of the 

Eighth IEEE International Conference on Computer Vision, 
p. 508–515.


modeling and rendering. IEEE Transactions on Pattern Analysis 

and Machine Intelligence, 24:1140–1146.

Li, D., G. Zhang, Z. Wu, and L. Yi, 2010. An edge embedded marker- 

based watershed algorithm for high spatial resolution remote 
sensing image segmentation. IEEE Transactions on Image 

Processing, 19:2781–2787.

Lowe, D.G., 1999. Object recognition from local scale-invariant 

features, Proceedings of the Seventh IEEE International 


confidence. IEEE Transactions on Pattern Analysis and Machine 

Intelligence, 23:1351–1365.


reconstruction in wide baseline stereo. Proceedings of the 17th 

International Conference on Pattern Recognition, pp. 76–79.

Moravec, H.P., 1981. Rover visual obstacle avoidance. Proceedings of the 

7th International Joint Conference on Artificial Intelligence- 


corner detection. Proceedings of the European Conference on 

Shan, J. and C.K. Toth, 2008. Topographic Laser Ranging and 

Scanning: Principles and Processing. Taylor & Francis Group/ 

CRC Press, Boca Raton, Florida.


matching with occlusions. Proceedings of Computer Vision- 

ECCV 2000, pp. 100–114.

Smith, S.M., and J.M. Brady, 1997. SUSAN - A new approach to 

tlow level image processing. International Journal of Computer 

Vision, 23:45–78.

Stentoumis, C., I. Grammatikopoulos, I. Kalisperakis, and G. Karras, 

2014. On accurate dense stereo-matching using a local adaptive 

multi-cost approach. ISPRS Journal of Photogrammetry and 


belief propagation. IEEE Transactions on Pattern Analysis and 

Machine Intelligence, 25:87–90.

Torr, P.H.S., 1995. Motion Segmentation and Outlier Detection, 

University of Oxford, UK.


Watson, 2010. Modelling the appearance and behaviour of urban 


25–42.

Van Meerbergen, G., M. Vergauwen, M. Pollefeys, and L. Van Gool, 

2002. A hierarchical symmetric stereo algorithm using dynamic 

programming. International Journal of Computer Vision, 


Veksler, O., 2005. Stereo correspondence by dynamic programming 
on-a tree. Proceedings of the 2005 IEEE Computer Society 
Conference on Computer Vision and Pattern Recognition (CVPR’05), 

multiphoto matching, International Journal of Computer Vision, 


Wu, B., Y. Zhang, and Q. Zhu, 2012. Integrated point and edge 

matching with color-weighted correlation, hierarchical belief 

propagation, and occlusion handling. IEEE Transactions on 

Pattern Analysis and Machine Intelligence, 34:492–504.

hierarchical image matching method for wide-baseline images, 


matching with color-weighted correlation, hierarchical belief 

propagation, and occlusion handling. IEEE Transactions on 

Pattern Analysis and Machine Intelligence, 34:492–504.