ISPRS Journal of Photogrammetry and Remote Sensing 66 (2011) S52–S61

Contents lists available at SciVerse ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Fusion of camera images and laser scans for wide baseline 3D scene alignment in urban environments ☆

Michael Ying Yang [a,*], Yanpeng Cao [b], John McDonald [b]

[a] Department of Photogrammetry, Institute of Geodesy and Geoinformation, University of Bonn, Bonn, Germany
[b] Department of Computer Science, National University of Ireland, Maynooth, Ireland

ARTICLE INFO

Article history:
Available online 7 October 2011

Keywords:
Senor fusion
Terrestrial laser scan
Wide baseline alignment
Viewpoint invariant feature
Plane extraction
Feature extraction

ABSTRACT

In this paper we address the problem of automatic laser scan registration in urban environments. This represents a challenging problem for two major reasons. First, two individual laser scans might be captured at significantly changed viewpoints (wide baseline) and have very little overlap. Second, man-made buildings usually contain many structures of similar appearances. This will result in considerable aliasing in the matching process. By sensor fusion of laser data with camera images, we propose a novel improvement to the existing 2D feature techniques to enable automatic 3D alignment between two widely separated scans. The key idea consists of extracting dominant planar structures from 3D point clouds and then utilizing the recovered 3D geometry to improve the performance of 2D image feature for wide baseline matching. The resulting feature descriptors become more robust to camera viewpoint changes after the procedure of viewpoint normalization. Moreover, the viewpoint normalized 2D features provide robust local feature information including patch scale and dominant orientation for effective repetitive structure matching in man-made environments. Comprehensive experimental evaluations with real data demonstrate the potential of the proposed method for automatic wide baseline 3D scan alignment in urban environments.

1. Introduction

Terrestrial laser scanners have become increasingly used for the collection of highly detailed 3D modeling in urban environments (Böhm, 2005; Cornelis et al., 2008). To obtain the full coverage of a complex urban scene, it is required that several scanning data are captured at different viewpoints. Therefore we need to register such individual scans into one global reference frame. For the registration, the common practice involves the manual deployment of highly distinguishable artificial targets in the scene. These target objects can be easily linked to generate 3D-to-3D tie objects between two separated 3D scans. To avoid the use of such artificial markers, a number of automatic matching algorithms have been proposed. The most popular method is the iterative closest point (ICP) algorithm developed by Besl and McKay (1992). However, the ICP algorithm requires a good initialization in order to converge to a global minimum, which can not be guaranteed without tedious user intervention. Nowadays, most laser manufacturers equip a high-resolution digital camera inside the scanner for users to capture digital imagery while the 3D point clouds are collected, so as to generate photorealistic 3D object and scene models. It stands to reason making use of the well-established image processing algorithms in the field of computer vision to improve the registration of 3D terrestrial laser scans.

Given images directly linked to the 3D point clouds, the focus of this paper is to automatically align two significantly separated laser scans (3D scene registration) via the matching of their associated 2D image appearances (2D image matching). This is a challenging problem for two major reasons. First, individual laser scans and their associated images might be captured at two very different viewpoints (wide baseline). The appearance of the same building facade will change significantly. Using the conventional 2D image features (Bay et al., 2008; Tuytelaars and Van Gool, 2004; Lowe, 2004; Donoser and Bischof, 2006; Mikolajczyk and Schmid, 2004), it is difficult to establish correct matches between such wide baseline image pairs. Second, man-made buildings usually consist of many structures of similar appearances (e.g. windows, doors, bricks). Such repetitive structures will cause considerable aliasing problem in the matching process. Based on comparing local appearances, it may be possible to match a single window in first image with any window in the second one. Hence any solution to the above problem must be able to handle large viewpoint changes and be robust to aliasing within the urban scenes.

In this paper, we present a complementary framework in which the 3D sensoring technology is integrated within the area of com-

☆ The first two authors contributed equally to this paper.
* Corresponding author. Tel.: +49 228 732906.
E-mail address: michaelyangying@uni-bonn.de (M.Y. Yang).

0924-2716/$ - see front matter © 2011 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.isprsjprs.2011.09.004

puter vision. We make use of the 3D information captured by a laser scanner to improve the performance of 2D image matching, and then apply the improved 2D image features to enable automatic registration of two widely separated 3D urban scenes. The framework includes two major steps. In the first step, we connect the images captured by a hand-held camera to the 3D laser data (the fusion of laser data with camera images). We obtain corresponding feature points between the camera captured images and the laser provided image, thus we can connect the image pixels to the 3D points (pixel-to-point correspondences). Since the photos are captured from similar viewpoints of the laser scanner, the standard SIFT matching is suitable for this task. In the second step, we present a novel method for generating viewpoint invariant features and demonstrate its application to robust matching over widely separated views in urban environments. In this step, we propose an effective method to extract a number of dominant planes in the 3D laser point cloud and then use them to describe the 3D spatial layout of the scanned scene. The 2D image features can be normalized with respect to these recovered 3D planes to achieve viewpoints invariance. The individual patches on the original image, each corresponding to an identified 3D planar region, are rectified to form the front-parallel views of building facades. Viewpoint invariant features are then extracted on these rectified views to provide a basis for further matching. The pixel-to-pixel feature correspondences allow us to link two separated 3D laser scans and subsequently register them into a global reference. The major procedure of the proposed framework is schematically illustrated in Fig. 1.

1.1. Major contributions

In this paper, we present a complementary sensor fusion framework. The advantages of 3D sensoring technology (accurate range capturing) and 2D computer vision research (robust image matching) are fully integrated. We make use of the 3D information captured by a laser scanner to improve the performance of 2D image matching, and subsequently apply the improved 2D image features to enable automatic registration of two widely separated 3D scenes. The key idea of this complementary sensor fusion framework is demonstrate in Fig. 2. Compared with some previous approaches on combining 2D feature with 3D geometry (Wu et al., 2008; Koeser and Koch, 2007), our method extracted a number of dominant 3D planes to represent the 3D layout of an urban setting. It is demonstrated the resulting piece-wise planar 3D model offers more robustness to the errors occurred in the process of 3D scanning.

In experiments, we systematically evaluate the performance of the proposed 3D viewpoint normalization. The results demonstrate

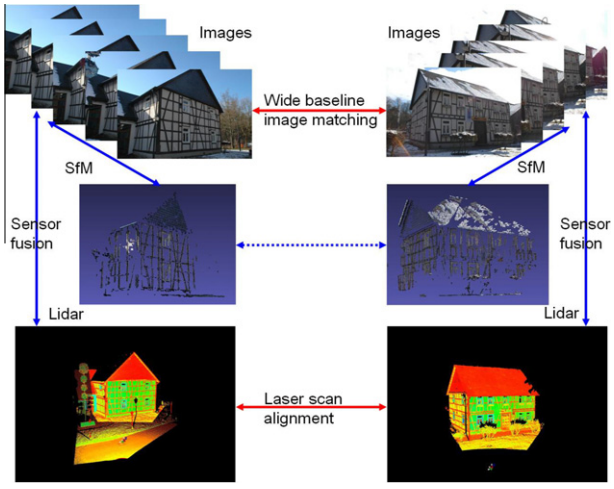


Fig. 1. The major procedure of the proposed framework.

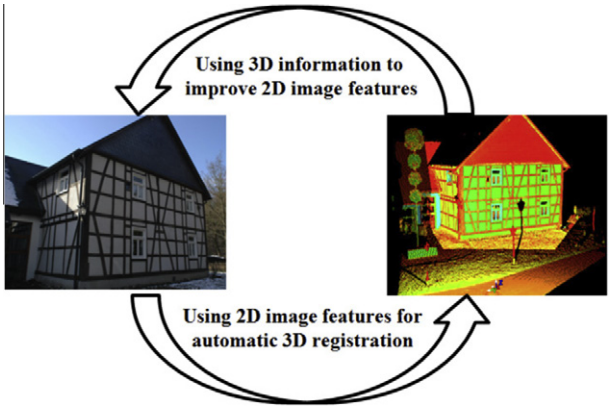


Fig. 2. The key idea of the proposed complementary sensor fusion framework. We make use of the 3D information captured by a laser scanner to improve the performance of 2D image matching, and subsequently apply the improved 2D image features to enable automatic registration of two widely separated 3D scenes.

that: (1) the resulting feature descriptors are very robust to the perspective distortions caused by large viewpoint changes, thus they are very suitable for wide baseline image matching; and (2) the features provide extra local feature information including patch scale and dominant orientation which can be used as effective geometric constraints between views. This makes viewpoint invariant features are robust to visual aliasing and suitable for repetitive structure matching in urban environments.

The remainder of the paper is organized as follows. Section 2 reviews some existing solutions for 3D model alignment and robust feature matching. In Section 3, we present the preprocessing step where the hand-held camera photos are linked to the 3D laser scans. In Section 4, we explain the procedures of 3D viewpoint normalization and propose an effective scheme to use the resulting viewpoint invariant features for repetitive structure matching in urban environments. In Section 5, the performance of the proposed method is comprehensively evaluated, and automatic alignment of wide baseline terrestrial laser scans is demonstrated as the application of this method. Finally, the conclusions are given in Section 6.

2. Related work

Terrestrial laser scanners are frequently used for the collection of highly detailed 3D urban modeling (Böhm, 2005; Cornelis et al., 2008). However, the complexity of natural scenes and the amount of information acquired by terrestrial laser scanners turn the registration among scans into a complex problem. Given two sets of 3D point clouds captured at different viewpoints, the task is to obtain tie points and to estimate an optimal transformation between them. Commercial softwares typically require users to manually deploy artificial targets as corresponding points between scenes. Recently, a number of matching algorithms have been proposed to avoid the manual intervention (Besl and McKay, 1992; Zhao et al., 2005; Pottmann et al., 2006). They are based on the iterative closest point (ICP) technique and to compute the alignment transformation by iteratively minimizing the sum of distances between closest points. Several improvements to the original ICP algorithm have also been proposed, such as the iterative closest points using invariant features (Sharp et al., 2002). However, the performances of ICP-based methods rely on a good estimation initialization and require good spatial configuration of 3D points which are not always guaranteed in realistic practices.

Since laser-scanning instruments nowadays are often equipped with an additional image sensor, many researchers proposed to enhance the performances of 3D point cloud alignment by referring to their associated 2D images. In (Seo et al., 2005), an effective method

was presented for automatic 3D model alignment via 2D image matching. Liu et al. (2006) presented a general framework to align 3D points from structure-from-motion (SfM) with range data. Images are linked to the 3D model to produce common points between range data. Ikeuchi et al. (2007) presented an automated 3D range to 3D range registration method that relies on the matching of reflectance range image and camera image. In (Gonzalez Aguilera et al., 2009), a flexible approach was presented for the automatic co-registration of terrestrial laser scanners and digital cameras by matching the camera images against the range image. Barnea and Filin (2007) presented a registration scheme to match the extracted features with 2D optical images using the scale invariant feature transform (SIFT) (Lowe, 2004). Similar to this method, Barnea and Filin (2008) developed a key-point based autonomous registration method using range images that also uses the 3D Euclidean distance between key-points as matched entities to identify correspondence. In Kang (2008) an automatic algorithm is described for registering terrestrial laser-scanning point clouds using reflectance images and SIFT features. However, these techniques only work well for frames with small observation changes (e.g. continuous videos) where the conventional feature techniques (e.g. Lowe, 2004; Donoser and Bischof, 2006) can produce robust image matching results. To produce satisfactory registration results of 3D points clouds captured at significantly changed viewpoints, we need a more effective image feature scheme which is capable of establishing robust correspondences between wide baseline image pairs.

Recently, some researchers in the field of computer vision have considered the use of 3D geometry as an additional cue to improve 2D feature matching. A novel feature detection scheme, viewpoint invariant patches (VIP), based on 3D normalized patches was proposed for 3D model matching and querying (Wu et al., 2008). In Koeser and Koch (2007), both texture and depth information were exploited for computing a normal view onto the surface. In this way they kept the descriptiveness of similarity invariant features (e.g. SIFT) while achieving extra invariance against perspective distortions. However the drawbacks of these methods is that they directly make use of the preliminary 3D point clouds from SfM. Viewpoint normalization with respect to the local computed tangent planes are prone to errors occurred in the process of 3D reconstruction. For predominantly planar scenes (urban environment), a piece-wise planar 3D model is more robust, compact, and efficient for viewpoint normalization of cameras with wide baselines (Yang et al., 2010).

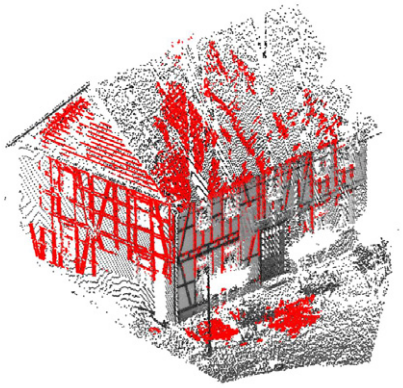
This paper further extends our previous work (Yang et al., 2010; Cao et al., 2011). We present a comprehensive set of experiments to evaluate the performance of the proposed viewpoint invariant features. After 3D viewpoint normalization, the resulting descriptors will remain less sensitive to viewpoint changes. Moreover, its shown that the features can provide robust information including patch scale and dominant orientation which in turn can be used as effective geometric constraints. Based on this fact, we further propose an effective framework for using viewpoint invariant features for challenging wide baseline matching tasks in urban environments. Here we accept multiple matches to cope with repetitive urban structures and then make use of the information associated with the extracted viewpoint invariant features (i.e. patch scale, dominant orientation, feature coordinates) to identify correct ones.

3. Preprocessing

Laser-scanning instruments nowadays are mostly equipped with an additional image sensor to capture digital imagery while the 3D range data were collected, so as to generate photorealistic 3D scene models. The captured image is directly linked to the 3D point cloud if the camera is directly integrated to the laser scanner. However, fixing the relative position between the 3D range and 2D image sensors has the following major limitations (Liu et al., 2006): (1) the acquisition of the images and range scans has to occur at the same viewpoint. This leads to a lack of 2D sensing flexibility since the limitations of 3D range sensor positioning such as stand-off distance and maximum distance; (2) the static arrangement of 3D and 2D sensors prevents the camera from being dynamically adjusted to the requirements of each particular scene; (3) sometimes users might need to capture images at different times, particularly if there were poor lighting conditions at the time that the range scans were acquired.

To overcome above drawbacks, we used an independent digital camera for image capturing and applied the technique described in

(a) 3D laser scan and a sequence of 2D images of the scanned scene.



(b) SfM point cloud mapping on laser scan.

Fig. 3. SfM point cloud mapping on laser scan. The red color points refer to 3D SfM from images. The density of laser scan is 1/100 of original data by resampling. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Liu et al. (2006) to link the camera images to the 3D range data. First, we capture a 3D laser scan and a sequence of 2D images of the scanned scene (see Fig. 3(a)). Note the images are gathered using an independent hand-held camera from various positions that do not necessarily coincide with the viewpoint of the range scanner. A sparse 3D point cloud can be reconstructed from these multiple-view images by using the standard SfM algorithm (Pollefeys et al., 2004). Then, the SIFT features are extracted on the camera images and the image captured by the laser for correspondence matching. A number of putative matches are found using local appearance descriptors, and then the RANdom SAmple Consensus (RANSAC) algorithm (Fischler and Bolles, 1981) is used to eliminate false correspondences by imposing a plane-to-plane mapping homography function. Since the viewpoint difference between the camera and the laser scanner is not significant, the standard SIFT technique is capable of producing robust image matches as tie points between the 3D SfM point cloud and the 3D laser scan. Finally, we compute the transformation that aligns the 3D models gathered via range sensing and computed via structure from motion, thus the complete set of 2D images is automatically linked to the 3D point cloud. Fig. 3(b) shows a result of such preprocessing.

The output of this preprocessing step is a number of 2D images linked to a 3D point cloud model (laser scan). Some previous works have proposed to directly extract SIFT features on these 2D images for 3D registration (Barnea and Filin, 2007; Kang, 2008; Barnea and Filin, 2008). In this paper, we further extend the idea of sensor fusion, making use of the recovered 3D information to improve the performance of 2D image feature extraction and matching.

4. 3D viewpoint invariant features

In this step, we present an effective method to extract a number of dominant planes in the 3D point cloud, as described in Yang and Förstner (2010). The detected 3D planes will be used to represent the spatial layout of an urban environment. The 2D image features are normalized with respect to these recovered planes to achieve better viewpoints invariance. A number of viewpoint invariant features are extracted on the rectified front-parallel views to provide a basis for further matching.

4.1. Dominant planes extraction

One of the most widely known methodologies for plane extraction is RANSAC algorithm. It has been proven to successfully detect planes in 2D as well as 3D. RANSAC is reliable even in the presence of a high proportion of outliers. Based on the observation that RANSAC may find wrong planes if the data has a complex geometry, we introduce a plane extraction method by integrating RANSAC and minimum description length (MDL). The MDL principle (Rissanen, 1978) is as follows. For a given data that consists of observations of some phenomenon, a number of different models is hypothesized to explain this phenomenon. Given a set of points, we assume several competing hypothesis, here namely, outliers (O), 1 plane and outliers (1P + O), 2 planes and outliers (2P + O), 3 planes and outliers (3P + O), 4 planes and outliers (4P + O), 5 planes and outliers (5P + O), etc.

Let n_0 points be given in a 3D coordinate and the coordinates be given up to a resolution of ϵ and be within range R. The description length for the n_0 points, when assuming outliers (O), therefore is

$$\#bits(points|\mathrm{O}) = n_0 \cdot (3lb(R/\epsilon)) \quad (1)$$

where $lb(R/\epsilon)$ bits are necessary to describe one coordinate.

If we now assume n points to sit on a plane and the other $\bar{n} = n_0 - n$ points to be outliers, we need

$$\#bits(points|1\mathrm{P}+\mathrm{O}) = n_0 + \bar{n}\cdot 3lb(R/\epsilon) + 3lb(R/\epsilon) + n\cdot 2lb(R/\epsilon) + \left[\sum_{i=1}^{n}\left\{\frac{1}{2ln2}\cdot(\boldsymbol{v}_i)^T\Sigma^{-1}(\boldsymbol{v}_i) + \frac{1}{2}lb(|\Sigma|/\epsilon^6) + \frac{3}{2}lb(2\pi)\right\}\right] \quad (2)$$

where the first term represents the n_0 bits for specifying whether a point is good or bad, the second term is the number of bits to describe the bad points, the third term is the number of bits to describe the 3 parameters of the plane, which is the number of bits to describe the model complexity, a variation of Rissanen (1978). We assumed the good points to randomly sit on the plane which leads to the fourth term, and to have Gaussian distributed derivations $\boldsymbol{v}_i$ from the plane with covariance matrix Σ. We show in Yang and Förstner (2010) that $\frac{1}{2ln2}\cdot(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-(\boldsymbol{\mu}) + \frac{1}{2}lb(|\Sigma|/\epsilon^6) + \frac{k}{2}lb(2\pi)$ bits are necessary to describe a Gaussian variable $\boldsymbol{x} \sim N(\boldsymbol{\mu},\Sigma)$, where $\boldsymbol{x} = (x_1,\ldots,x_k)^T$, $\boldsymbol{\mu}$ and Σ are given, and if it is rounded to multiples of ϵ.

$\#bits(points{-}2\mathrm{P}+\mathrm{O})$, $\#bits(points{-}3\mathrm{P}+\mathrm{O})$, $\#bits(points{-}4\mathrm{P}+\mathrm{O})$, and $\#bits(points{-}5\mathrm{P}+\mathrm{O})$, and so on, can be deducted in a similar way (Yang and Förstner, 2010).

Incremental RANSAC is applied to extract planes in the point cloud. The MDL principle, deducted above, for interpreting a set of points in 3D space,is employed to decide which hypothesis is the best one. This method of integrating RANSAC and MDL has been shown to avoid detecting wrong planes (Yang and Förstner, 2010). One example demonstrating dominant plane extraction is shown in Fig. 10.

4.2. Viewpoint invariant feature generation

In this step, we perform normalization with respect to the extracted dominant 3D planes to achieve viewpoint invariance. Given a perspective image of a world plane, the goal is to generate the front-parallel view of the plane. This is equivalent to obtaining the image of a world plane where the camera viewing direction is parallel to the plane normal. It is well known that the mapping between a 3D world plane and its perspective image is defined as a 3×3 homography. Since we know the 3D positions of the points on the building facade and their corresponding image coordinates, we can compute the homography relating the facade plane to its image given at least four correspondences. The computed homography H enables us to warp the original image to a normalized front-parallel view where the perspective distortion is removed. Fig. 4 shows some results of such viewpoint normalization.

Within the normalized front-parallel views of the scene, the viewpoint invariant features are computed in the same manner as the SIFT scheme (Lowe, 2004). Given a number of extracted dominant 3D planes, features extraction can be efficiently performed in a single pass w.r.t. the planes. Potential keypoints are identified by scanning local extreme in a series of difference-of-Gaussian (DoG) images. For each detected keypoint, appropriate scale and orientation are assigned to it and a 128-element SIFT descriptor is created based upon image gradients of its local neighborhood. A complete viewpoint invariant feature consists of the following components: (1) $\boldsymbol{X}$ is its 3D position in the space; (2) $\boldsymbol{x}$ is its 2D coordinates in the normalized front-parallel view; (3) s is its corresponding spatial patch scale; (4) θ is the dominant gradient orientation of the normalized patch; (5) $\boldsymbol{f}$ is the 128-element descriptor; and (6) $\boldsymbol{n}$ is the normal of the plane it belongs to.

4.3. Repetitive structure matching in urban environments

In this section we propose an effective framework for robust wide baseline image matching in urban environments using the extracted viewpoint invariant features. We follow the commonly used image matching scheme of: (1) establishing a set of putative correspondences based on the matching local descriptors, and (2)

Fig. 4. Some examples of viewpoint normalization. *Left*: original images. *Right*: normalized front views. Note the perspective distortions are largely reduced in the warped front-parallel views of the building walls (e.g. a rectangular window in the 3D world will also appear rectangular in the normalized images).

computing a global geometric constraint to identify true correspondences across the views.

Given a number of extracted viewpoint invariant features, we applied the criterion described in Zhang and Košecká (2007) to generate the putative correspondences. We consider two features matched if the cosine of the angle between their associated descriptors $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ is above some threshold δ as:

$$\cos(\boldsymbol{f}_i, \boldsymbol{f}_j) = \frac{\boldsymbol{f}_i \cdot \boldsymbol{f}_j}{\|\boldsymbol{f}_i\|_2 \|\boldsymbol{f}_j\|_2} > \delta \tag{3}$$

where $\|\cdot\|_2$ represents the L_2-norm of a vector. The essence of this criterion is to establish matches between features having similar descriptors. In cases where multiple matches meet the criteria, we accept the top 10 matches. In urban environments where many repetitive structures (e.g. windows) exist, this criterion establishes matches between features having similar descriptors. This keeps the potential correspondences extracted on the images of repetitive structures for further geometric verification. However, Eq. (3) is a quite loose criterion. The resulting putative set will contain a large percentage of outliers (90–95%), within which we need to effectively identify the correct correspondences.

After establishing a set of putative feature matches based on the matching of local descriptors, we need to refine the results and to identify the true correspondences by imposing a geometric constraint. The RANSAC technique (Fischler and Bolles, 1981) is usually applied for this task. In RANSAC iterations, the correspondences consistent with each generated hypothesis (e.g. the symmetric transfer error is less than a threshold) are defined as its inliers. The hypothesis with the most supports is chosen and its corresponding inliers are defined as true matches. The number of samples M required to guarantee a confidence ρ that at least one sample is outlier free is computed as:

$$M = \frac{ln(1-\rho)}{ln(1-(1-\eta)^P)} \tag{4}$$

where η is the percentage of outliers and P is the number of observations required to generate a hypothesis per sample. It is noted M is dependent on both the complexity of the imposed geometric model and the fraction of outliers. Using the conventional 2D feature schemes (e.g. SIFT or MSER), only the 2D image coordinates of SIFT features can be used to generate geometric constraints (e.g. F-Matrix or H-Matrix). Therefore, a number of SIFT feature matches are required to compute the F-Matrix (7 correspondences) or the H-matrix (4 correspondences). When the fraction of outliers is significant, RANSAC needs a large number of samples using a complex geometric model.

In comparison, the viewpoint invariant features are extracted on the front-parallel views of the building facade, taken at different distances and up to a camera translation and rotation around its optical axis. Every feature correspondence provides three constraints: scale (camera distance), 2D coordinates on the canonical view (camera translation), and dominant orientation (rotation around its optical axis). Therefore, a single feature correspondence is enough to completely define a homothetic mapping relation between two canonical views. Consider a pair of matched features $(\boldsymbol{x}_1^m, s_1^m, \theta_1^m)$ and $(\boldsymbol{x}_2^n, s_2^n, \theta_2^n)$ both extracted on the normalized front-parallel views, a geometric constraint between two views is generated as follows:

$$\begin{bmatrix} x_1 - x_1^m \\ y_1 - y_1^m \\ 1 \end{bmatrix} = \begin{bmatrix} \Delta s & 0 & 0 \\ 0 & \Delta s & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\Delta\theta & -\sin\Delta\theta & 0 \\ \sin\Delta\theta & \cos\Delta\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_2 - x_2^m \\ y_2 - y_2^m \\ 1 \end{bmatrix} \tag{5}$$

where $\Delta s = s_1^m / s_2^n$ is the scale ratio and $\Delta\theta = \theta_1^m - \theta_2^n$ is the orientation difference. Using this simplified geometric model, a much smaller number of samples are needed to guarantee the generation of the correct hypothesis. The comparative result is shown in Table 1. It is noted that the required number for RANSAC sampling decreased significantly due to the use of the simplified geometric model. Moreover, RANSAC can successfully return the true correspondences from a putative feature set of high outlier percentage. As shown in Table 1 , the true correspondences can be identified from a putative set containing 95% outliers within a few iterations. Therefore we can set a weak criteria Eq. (3) to establish a large number of putative matches (i.e. containing a large number of outliers) and in this way deal with high levels of aliasing in the scene. In urban environments where lots of respective structures (e.g. windows, doors, bricks) exist, setting a strict matching criteria (e.g. ratio check (Lowe, 2004)) will initially sacrifice many true correspondences.

In next section, our experimental evaluations will show that for all ground true correspondences the scale ratios and orientation differences are equal up to a very small offset. It means that the information of patch scale and dominant orientation associated with the viewpoint invariant features are robust enough to generate geometrical hypothesis, which is impossible in the conventional 2D feature schemes.

5. Experimental results

In this section, we conducted experiments to evaluate the performance of the proposed viewpoint invariant features and demonstrated their applications for automatic alignment of wide baseline terrestrial laser scans, with focus in the urban environments. Two groups of experiments were organized to demonstrate the advantages of such complementary sensor fusion. The first experiment shows that the performances of 2D image feature matching can be significantly improved after taking into account

Table 1
The theoretical number of samples (M) required for RANSAC to ensure 95% confidence (ρ) that one outlier free sample is obtained for estimation of geometrical constraint. The actual required number is around an order of magnitude more.

Outlier percentage	40%	60%	80%	90%	95%
Our method (1 point)	4	6	14	29	59
H-matrix (4 point)	22	116	1871	2.996×10^4	4.793×10^5
F-matrix (7 point)	106	1827	2.340×10^5	2.996×10^7	3.835×10^9

the underlying 3D geometry. Then two widely separated 3D scenes can be robustly aligned using the improved 2D image features (viewpoint invariant feature), as shown in the second experiment.

5.1. Performance evaluations

The first experiment is to demonstrate that the performances of 2D image feature matching can be significantly improved by taking into account the underlying 3D structure. We collected 10 pairs of images over largely separated views with a calibrated camera. Each pair consists of 10 images, of which 5 images from the left view, the other 5 images from the right view. Then, we applied orientation software AURELO (Läbe and Förstner, 2006) to achieve full automatic relative orientation of these multi-view images. Then we used software PMVS (patch-based multi-view stereo) (Furukawa and Ponce, 2010) for deriving a dense point cloud for each view of image pairs. It provides a set of 3D points with normals at those positions where there is enough texture in the images. In total, 20 SfM 3D point clouds are generated which covered 10 general urban scenes from two significantly changed viewpoints. Some representative results are given in Fig. 5.

For each point cloud, a number of dominant planes were extracted to represent the underlying 3D structures. Given the extracted dominant planes, we perform normalization w.r.t. these planes to achieve viewpoint invariance. After viewpoint normalization, corresponding scene elements will have more similar appearances. The resulting features will suffer less from the perspective distortions and show better descriptiveness. We tested our method on 10 pairs of wide baseline 3D point clouds, as shown in Fig. 5, to demonstrate such improvements. It is noted that both 3D point clouds covered a same dominant planar structure which can be easily related through a homography. A number of SIFT and viewpoint invariant features were extracted on the original images and on the normalized front-parallel views, respectively. Then we followed the method described in Mikolajczyk and Schmid (2005) to define a set of ground truth matches. The extracted features in the first image were projected onto the second one using the homography relating the images (we manually selected 4 well conditioned correspondences to calculate the homography). A pair of features is considered matched if the overlap error of their corresponding regions is minimal and less than a threshold (Mikolajczyk and Schmid, 2005). We adjusted the threshold value to vary the number of resulting feature correspondences.

We quantitatively measured how well two correctly matched features relate with each other in terms of the Euclidean distance between their corresponding descriptors, their scale ratio, and their orientation difference. For each point cloud pair, we select 200 correspondences and calculate the average Euclidean distance between their descriptors. The quantitative results are shown in Fig. 6. It is noted that the procedure of viewpoint normalization will compensate the effects of perspective distortion. The Euclidean distance between the matched features decreased significantly from 0.6751 (the average for 10 image pairs) to 0.4067 after the procedure of viewpoint normalization. It means the resulting feature descriptors remain less sensitive when viewpoint changes. For each pair of matched features, we also computed the difference between their dominant orientations and the ratio between their patch scales. The results are shown in Figs. 7 and 8, respectively. On the normalized parallel-front views, the viewing direction is normal to the extracted 3D plane and the camera roll angle becomes zero. The matched features extracted on such normalized views should have the same dominant orientations and scale ratios. In experiments, it is observed that the dominant orientations and scale ratios are equal up to a very small tolerance for all true correspondences after viewpoint normalization.

To qualitatively demonstrate the improvements, we have chosen a pair of wide baseline point clouds and shown a number of

Fig. 5. Four pairs of 3D point clouds and their associated images captured at widely separated views.

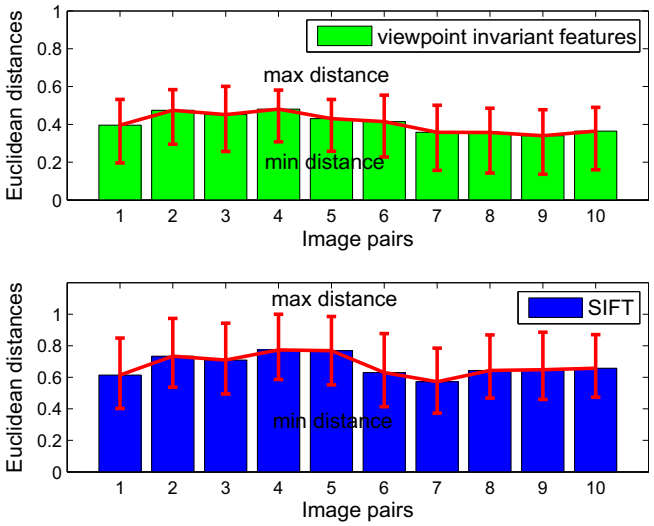


Fig. 6. Performance comparison between SIFT and viewpoint invariant features: the average Euclidean distances between the descriptors of matched features.

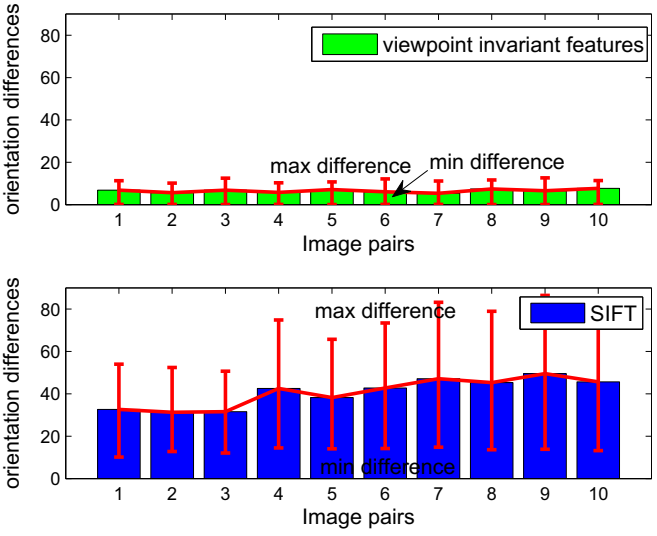


Fig. 7. Performance comparison between SIFT and viewpoint invariant features: the orientation differences between matched features.

matched features on the original images (cf. Fig. 9, left) and on the normalized images (Fig. 9, right). Their corresponding scales and orientations are also displayed. On the front-parallel views, the matched features have very similar orientations. Also their scale ratios show better consistency. The results demonstrated that we can robustly make use of the scale and orientation information associated with local image features to generate geometrical con-

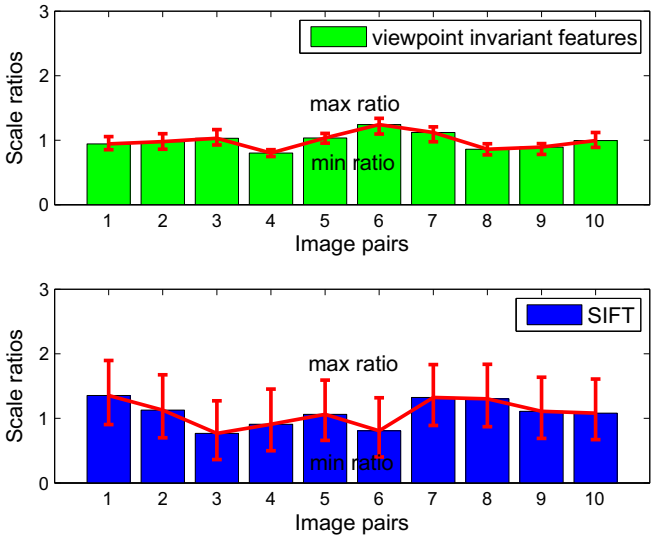


Fig. 8. Performance comparison between SIFT and viewpoint invariant features: the scale ratios between matched feature. The matched feature extracted on the normalized front-parallel views show better robustness to viewpoint changes.

straints between images. For the viewpoint invariant features, a single correspondence is enough to completely determine a homothetic mapping relation (Eq. (5)). Using this simplified model, a much smaller number of samples are required to guarantee the generation of a correct hypothesis in the RANSAC iterations.

5.2. Wide baseline laser scan alignment

In this section, we apply the proposed viewpoint normalized features to automatically align laser scans captured at widely separated viewpoints. We have taken two groups of laser scanning data using Leica HDS6000. Each group contains two individual 3D point clouds of a same building captured at largely separated viewpoints. The laser-equipped camera simultaneously captured an intensity image which provides pixel-to-point correspondences to the 3D point cloud. For each laser scan, we also took 5 images using a hand-held camera at similar viewpoints.

For the preprocessing step, we applied the standard SIFT scheme (Lowe, 2004) to match the 2D camera images to the laser-provided image, as described in Section 3. Since the viewpoint change between camera and laser is not significant, the SIFT technique can produce robust image matches as tie points between the camera images and the 3D laser scan. The pixel-to-pixel tie points allow us to link the 3D point cloud from SfM to the laser scanning (cf. an example shown in Fig. 3). Finally, a number of dominant planes were extracted from each point cloud, while the remaining 3D points were removed. Fig. 10 shows an example of the captured laser scan and the extracted dominant 3D planes. Fig. 11 shows the histograms of corresponding 3D points to the closest extracted dominant plane. As seen in this image, noisy points are comparable

Fig. 9. A number of matched features are shown. *Left*: on the original images. *Right*: on the front parallel views. Their scales and orientations are annotated. The feature matches on the viewpoint normalized views have very similar orientations and consistent scale ratios.

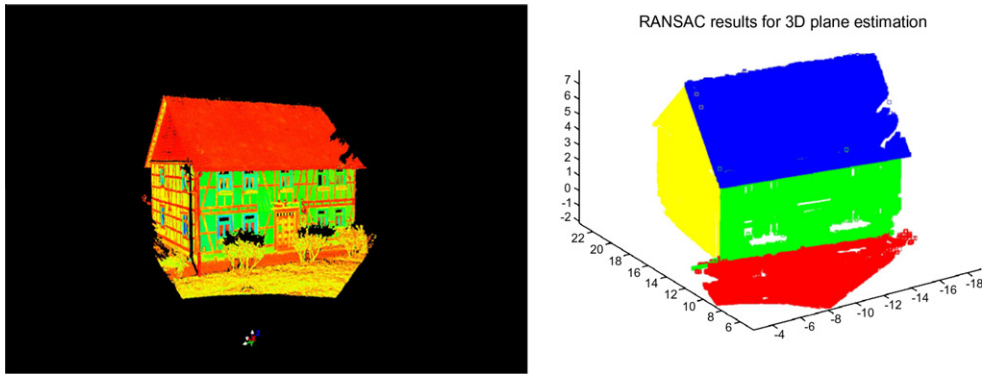


Fig. 10. *Left*: a snap-shot image of 3D point cloud taken by laser scanner. *Right*: the four dominant planes automatically extracted from the point cloud.

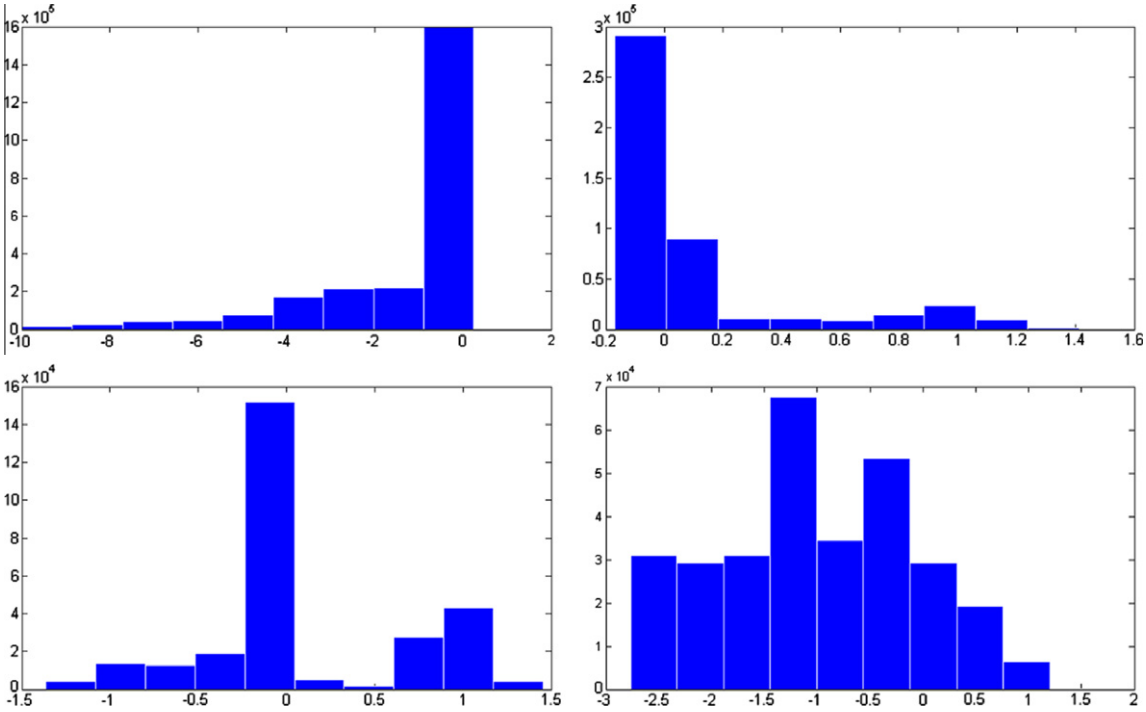


Fig. 11. The histograms of corresponding 3D points to the closest extracted dominant plane in Fig. 10. Noisy points are comparable to points supporting the dominant plane, or even more than the points supporting the dominant plane.

to points supporting the dominant plane (Fig. 11, bottom left), or even more than the points supporting the dominant plane (Fig. 11, bottom right). Therefore, tangent-plane based approach used in Wu et al. (2008) does not work well here.

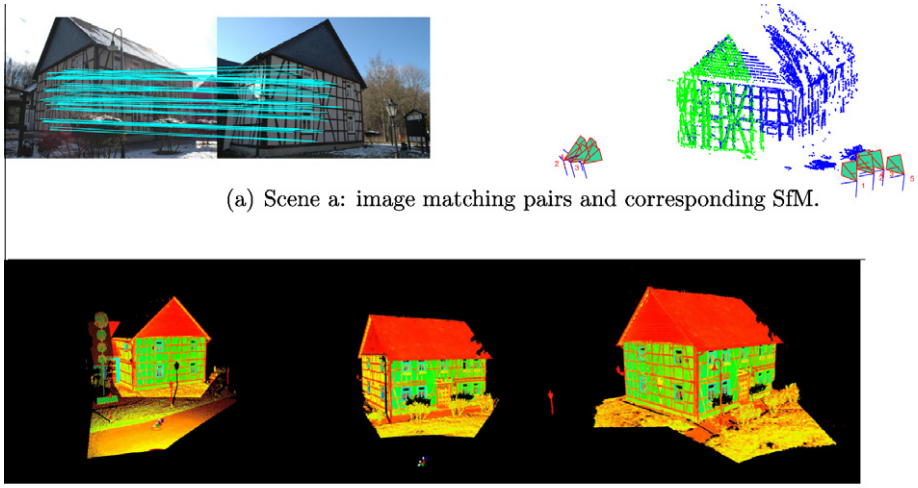
Then, we applied the proposed framework described in Section 4.3 for the difficult wide baseline matching tasks. A set of putative matches were firstly established based on the criterion described in Eq. (3), among them the inlier correspondences were selected by imposing the geometrical constraint Eq. (5). For comparison, we applied the scale-invariant feature scheme SIFT and the affine-invariant feature scheme MSER for the same task. A set of putative matches were firstly established. In this step, we need to set a strict criterion to make the resulting putative sets have a good portion of inliers (more than 20%), otherwise RANSAC needs a large number of iterations to return the correct correspondences (refer to Table 1). In experiments, we applied the ratio check scheme (Lowe, 2004) and set the ratio threshold at 0.85. Setting a strict matching criterion (ratio check (Lowe, 2004)) will initially sacrifice many true correspondences. Then we used RANSAC to compute the correct H-matrix to identify inlier correspondences.

The matching results are shown in Figs. 12 and 13 with the quantitative comparisons provided in Table 2. It is noted that the viewpoint invariant features can handle the large viewpoint changes (the view angles changed more than 90 degrees), for which SIFT and MSER do not work well. Moreover, using the proposed framework in Section 4.3, we can establish many correct feature correspondences in the presence of substantial repetitive structures (e.g. windows, bricks).

Finally, we computed the transform matrix relating two individual laser scans given a number of matched 3D points. The laser alignment results are shown in Figs. 12 and 13.

5.3. Summary

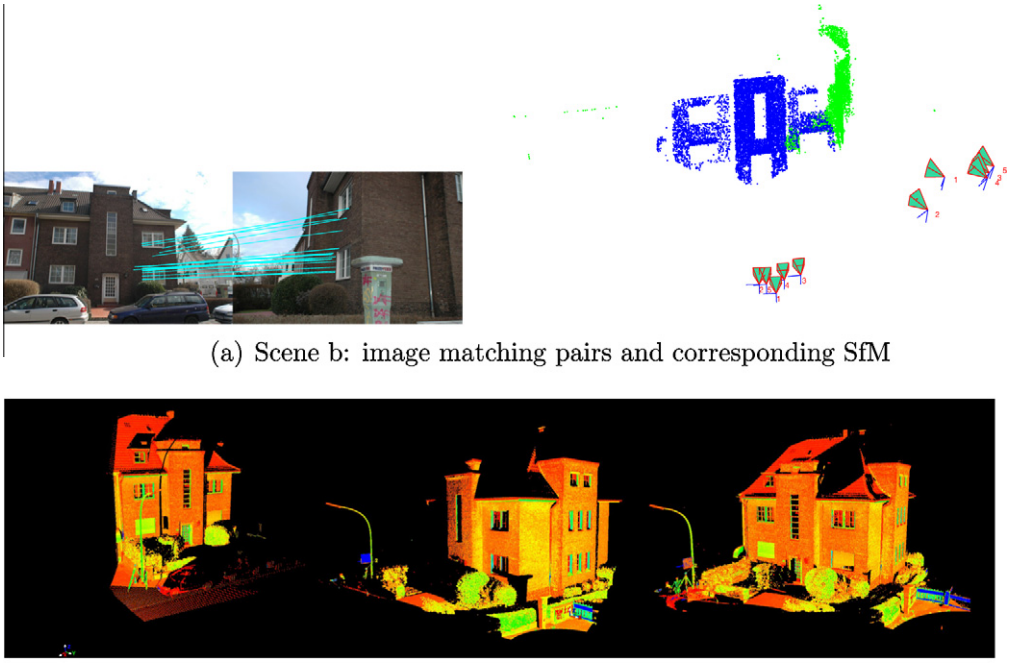
In terms of wide baseline matching in urban environments, the proposed viewpoint invariant feature achieves a twofold improvement. First, the procedure of viewpoint normalization ensures that



(a) Scene a: image matching pairs and corresponding SfM.

(b) Two different views of laser scan and corresponding alignment result.

Fig. 12. One example of wide baseline 3D scene matching. Significant viewpoint changes can be observed on the associated image pairs. Moreover, the building facades contains many structures of similar appearances (e.g. windows).



(a) Scene b: image matching pairs and corresponding SfM

(b) Two different views of laser scan and corresponding alignment result.

Fig. 13. Another example of wide baseline 3D scene matching. Also, significant viewpoint changes can be observed on the associated image pairs and lots of repetitive structures exist.

Table 2
The quantitative results of wide baseline 3D scene matching. (I – the number of initial correspondences by matching descriptors, N – the number of inliers correspondences returned by the RANSAC technique, T – the number of correct ones.)

Scene	SIFT			MSER			Our method		
	T	N	I	T	N	I	T	N	I
Scene a	19	28	127	7	16	100	79	80	901
Scene b	0	13	53	3	13	57	24	25	658

even in situations of considerable perspective distortion the resulting feature descriptors remain less sensitive, as shown in Fig. 6. Using the improved local descriptors enables us to establish correct correspondences over widely separated images. Second, the scale and orientation information associated with viewpoint invariant features can be used for estimating simplified global geometric constraints that are robust to visual aliasing (see Figs. 7 and 8). This makes viewpoint invariant features particularly suitable for image matching in urban environments where lots of repetitive structures exist.

6. Conclusions

Nowadays most laser-scanning equipments are accompanied with an additional image camera. In this paper we have proposed an effective sensor fusion framework which consists of a laser

scanner and a hand-held digital camera. We make use of the 3D information captured by a laser scanner to improve the performance of 2D image matching, and then apply the improved 2D image features to enable automatic registration of two widely separated 3D scenes. To achieve this, we brought in the concept of 3D viewpoint normalization and extracted features on the normalized front-parallel views w.r.t. 3D dominant planes derived from the point cloud of a scene. The resulting viewpoint invariant features enable us to link the corresponding 3D points automatically in terms of wide baseline image matching. We evaluated the proposed feature matching scheme against the conventional 2D feature detectors, and applied it to realistic wide baseline laser scanning data of a variety of urban scenes. The experimental results demonstrate the potential of viewpoint invariant features for robust and automatic wide baseline laser scan registration.

Based on piece-wise planar scene assumption, our framework can automatically align two wide baseline laser scans, which share at least one dominant plane. But the proposed method will not produce reliable results in case the building facades are poorly textured. In the future, we will further extend the method for laser scans captured in more complex and larger scale environments. Instead of a piece-wise planar scene assumption, we will need a more general 3D object extraction algorithm to deal with irregular 3D building structures. Eventually the method will be used as an important component in applications such as user navigation, augmented reality, and intelligent robotics in urban environments.

Acknowledgments

The work is funded by Deutsche Forschungsgemeinschaft (German Research Foundation) FO 180/16-1, and a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge these supports.We would like to thank Thomas Läbe and Wolfgang Förstner for providing the AURELO software and the helpful suggestions. We thank Yasutaka Furukawa and Jean Ponce for providing the PMVS software. We would also like to thank the anonymous reviewers for their valuable comments and suggestions.

References

Barnea, S., Filin, S., 2007. Registration of terrestrial laser scans via image based features. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 36 (Part 3/W52), 32–37.

Barnea, S., Filin, S., 2008. Keypoint based autonomous registration of terrestrial laser point-clouds. ISPRS Journal of Photogrammetry and Remote Sensing 63 (1), 19–35.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Computer Vision and Image Understanding 110 (3), 346–359.

Besl, P., McKay, N., 1992. A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2), 239–256.

Böhm, J., 2005. Terrestrial laser scanning – a supplementary approach for 3d documentation and animation. In: Photogrammetric Week 2005, pp. 263–271.

Cao, Y., Yang, M.Y., McDonald, J., 2011. Robust alignment of wide baseline terrestrial laser scans via 3d viewpoint normalization. In: IEEE Workshop on Applications of Computer Vision, pp. 455–462.

Cornelis, N., Leibe, B., Cornelis, K., Gool, L., 2008. 3D urban scene modeling integrating recognition and reconstruction. International Journal of Computer Vision 78 (2–3), 121–141.

Donoser, M., Bischof, H., 2006. Efficient maximally stable extremal region (MSER) tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 553–560.

Fischler, M., Bolles, R., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24 (6), 381–395.

Furukawa, Y., Ponce, J., 2010. Accurate, dense, and robust multi-view stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (8), 1362–1376.

Gonzalez Aguilera, D., Rodriguez Gonzalvez, P., Gomez Lahoz, J., 2009. An automatic procedure for co-registration of terrestrial laser scanners and digital cameras. ISPRS Journal of Photogrammetry and Remote Sensing 64 (3), 308–316.

Ikeuchi, K., Oishi, T., Takamatsu, J., Sagawa, R., Nakazawa, A., Kurazume, R., Nishino, K., Kamakura, M., Okamoto, Y., 2007. The great buddha project: digitally archiving, restoring, and analyzing cultural heritage objects. International Journal of Computer Vision 75 (1), 189–208.

Kang, Z., 2008. Automatic registration of terrestrial point cloud using panoramic reflectance images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 37 (Part B5), 431–436.

Koeser, K., Koch, R., 2007. Perspectively invariant normal features. In: IEEE International Conference on Computer Vision, pp. 1–8.

Läbe, T., Förstner, W., 2006. Automatic relative orientation of images. In: Proceedings of the 5th Turkish–German Joint Geodetic Days, Berlin.

Liu, L., Stamos, I., Yu, G., Wolberg, G., Zokai, S., 2006. Multiview geometry for texture mapping 2d images onto 3d range data. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2293–2300.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2), 91–110.

Mikolajczyk, K., Schmid, C., 2004. Scale and affine invariant interest point detectors. International Journal of Computer Vision 60 (1), 63–86.

Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (10), 1615–1630.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R., 2004. Visual modeling with a hand-held camera. International Journal of Computer Vision 59 (3), 207–232.

Pottmann, H., Huang, Q., Yang, Y., Hu, S., 2006. Geometry and convergence analysis of algorithms for registration of 3d shapes. International Journal of Computer Vision 67 (3), 277–296.

Rissanen, J., 1978. Modelling by shortest data description. Automatica 14 (5), 465–471.

Seo, J., Sharp, G., Lee, S., 2005. Range data registration using photometric features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1140–1145.

Sharp, G.C., Lee, S.W., Wehe, D.K., 2002. Icp registration using invariant features. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (1), 90–102.

Tuytelaars, T., Van Gool, L., 2004. Matching widely separated views based on affine invariant regions. International Journal of Computer Vision 59 (1), 61–85.

Wu, C., Clipp, B., Li, X., Frahm, J., Pollefeys, M., 2008. 3D model matching with viewpoint-invariant patches (VIP). In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Yang, M.Y., Cao, Y., Förstner, W., McDonald, J., 2010. Robust wide baseline scene alignment based on 3d viewpoint normalization. In: International Symposium on Visual Computing, pp. 654–665.

Yang, M.Y., Förstner, W., 2010. Plane detection in point cloud data. Tech. Rep. TR-IGG-P-2010-01, Department of Photogrammetry, Institute of Geodesy and Geoinformation, University of Bonn.

Zhang, W., Košecká, J., 2007. Hierarchical building recognition. Image and Vision Computing 25 (5), 704–716.

Zhao, W., Nister, D., Hsu, S., 2005. Alignment of continuous video onto 3d point clouds. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8), 1305–1318.