FINDING A GOOD FEATURE DETECTOR-DESCRIPTOR COMBINATION FOR THE 2D KEYPOINT-BASED REGISTRATION OF TLS POINT CLOUDS

S. Urban, M. Weinmann

Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology (KIT)
Englerstr. 7, 76131 Karlsruhe, Germany - {steffen.urban, martin.weinmann}@kit.edu

Commission III, WG III/2

KEY WORDS: Laser scanning, TLS, point cloud, imagery, feature extraction, local features, registration

ABSTRACT:

The automatic and accurate registration of terrestrial laser scanning (TLS) data is a topic of great interest in the domains of city modeling, construction surveying or cultural heritage. While numerous of the most recent approaches focus on keypoint-based point cloud registration relying on forward-projected 2D keypoints detected in panoramic intensity images, little attention has been paid to the selection of appropriate keypoint detector-descriptor combinations. Instead, keypoints are commonly detected and described by applying well-known methods such as the Scale Invariant Feature Transform (SIFT) or Speeded-Up Robust Features (SURF). In this paper, we present a framework for evaluating the influence of different keypoint detector-descriptor combinations on the results of point cloud registration. For this purpose, we involve five different approaches for extracting local features from the panoramic intensity images and exploit the range information of putative feature correspondences in order to define bearing vectors which, in turn, may be exploited to transfer the task of point cloud registration from the object space to the observation space. With an extensive evaluation of our framework on a standard benchmark TLS dataset, we clearly demonstrate that replacing SIFT and SURF detectors and descriptors by more recent approaches significantly alleviates point cloud registration in terms of accuracy, efficiency and robustness.

1 INTRODUCTION

In order to obtain a highly accurate and detailed acquisition of local 3D object surfaces within outdoor environments, terrestrial laser scanning (TLS) systems are used for a variety of applications in the domains of city modeling, construction surveying or cultural heritage. Each captured TLS scan is represented in the form of a point cloud consisting of a large number of scanned 3D points and, optionally, additional attributes for each point such as intensity information. In order to provide a dense and (almost) complete 3D acquisition of interesting parts of a scene, typically, multiple scans have to be captured from different locations and – since the spatial 3D information is only measured w.r.t. the local coordinate frame of the laser scanner – it is desirable to automatically estimate the respective transformations in order to align all captured scans in a common coordinate frame. The estimation of these transformations is commonly referred to as point cloud registration, and nowadays most approaches rely on specific features extracted from the point clouds instead of using the complete point clouds.

Despite a variety of features which may be extracted from point clouds and alleviate point cloud registration (e.g. planes or lines), we focus on the use of keypoints, since these may efficiently be extracted as local features. While approaches for detecting and describing 3D keypoints have recently been involved for point cloud registration (Theiler et al., 2013; Theiler et al., 2014), such a strategy typically relies on a subsampling of the original point (e.g. via voxel grid filtering) in order to get an approximately homogeneous point density. The alternative of directly working on the captured data consists of exploiting the discrete (spherical or cylindrical) scan pattern and deriving 2D image representations for either range or intensity information. Particularly in the intensity images, distinctive 2D keypoints may efficiently be derived and – in case they are part of any feature correspondence between different images – subsequently projected to 3D space by exploiting the corresponding range information which, in turn, yields sparse point sets of corresponding 3D points.

Concerning the use of 2D keypoints, those keypoint detectors and descriptors presented with the Scale Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) are typically exploited. In the absence of noise, a low number of feature correspondences may already be sufficient to obtain a suitable estimate for the relative orientation between two 3D point sets. However, the range measurements of a TLS system are typically corrupted with a certain amount of noise which requires additional effort. Consequently, it has been proposed to increase the reliability of the estimate by assigning each keypoint a quality measure which indicates the reliability of the respective range information and allows discarding unreliable keypoints (Barnea and Filin, 2008; Weinmann and Jutzi, 2011). Furthermore, the feature matching may result in a certain amount of wrong feature correspondences which may be identified by involving the well-known RANdom SAmple Consensus (RANSAC) algorithm (Fischler and Bolles, 1981) for robustly estimating a given transformation model (Barnea and Filin, 2007; Boehm and Becker, 2007). However, little attention has been paid to the fact that more recent approaches for extracting local features seem to overcome the main limitations of SIFT and SURF by increasing computational efficiency and simultaneously delivering even more feature correspondences which are still reliable and thus may significantly contribute to obtain a suitable estimate.

In this paper, we present a framework involving different modern 2D keypoints detectors and descriptors for finding corresponding image contents in the panoramic intensity images derived for the single scans. The respective forward-projection to 3D space yields sparse point sets of corresponding 3D points. Instead of directly exploiting these corresponding 3D points for point cloud registration, however, we exploit the bearing vectors defined by the origin of the local coordinate frame and the sparse 3D point sets, since these bearing vectors may be determined with a higher reliability in comparison to range measurements. This allows us to transfer the task of point cloud registration to the task of finding the relative orientation between sets of bearing vectors which may efficiently be handled by well-known algorithms for estimat-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
Editors: S. Oude Elberink, A. Velizhev, R. Lindenbergh, S. Kaasalainen, and F. Pirotti
doi:10.5194/isprsannals-II-3-W5-121-2015

ing the pose of omnivision cameras. As main contributions,

- we involve different approaches for extracting local features from the panoramic intensity images,
- we exploit the local features and the corresponding range information to define the respective bearing vectors,
- we transfer the task of point cloud registration from the object space (i.e. the direct alignment of 3D point sets) to the observation space (i.e. the direct alignment of sets of bearing vectors), and
- we investigate the influence of the feature extraction methods on the results of point cloud registration.

After briefly reviewing related work w.r.t. keypoint extraction and matching and w.r.t. keypoint-based point cloud registration in Section 2, we present the different methods involved in our framework in Section 3. Subsequently, in Section 4, we provide an extensive evaluation of our framework on a standard benchmark TLS dataset and discuss the derived results w.r.t. performance, efficiency and robustness. Finally, in Section 5, we draw conclusions and outline ideas for future research.

2 RELATED WORK

In our work, we focus on keypoint-based point cloud registration, where the keypoints are derived from 2D imagery. In the following, we hence reflect different approaches to detect and describe such keypoints representing the basis for deriving sparse 3D point sets (Section 3.1) and, subsequently, we discuss how the corresponding sparse 3D point sets may be aligned in a common coordinate frame (Section 2.2).

2.1 Keypoint Extraction and Matching

Generally, different types of visual features may be extracted from images in order to detect corresponding image contents (Weinmann, 2013). However, local features such as corners, blobs or small image regions offer significant advantages. Since such local features (*i*) may be extracted very efficiently, (*ii*) are accurately localized, (*iii*) remain stable over reasonably varying viewpoints and (*iv*) allow an individual identification, they are well-suited for a variety of applications such as object recognition, autonomous navigation and exploration, image and video retrieval, image registration or the reconstruction, interpretation and understanding of scenes (Tuytelaars and Mikolajczyk, 2008; Weinmann, 2013). Generally, the extraction of local features consists of two steps represented by feature detection and feature description.

For feature detection, corner detectors such as the Harris corner detector (Harris and Stephens, 1988) or the Features from Accelerated Segment Test (FAST) detector (Rosten and Drummond, 2005) are widely used. The detection of blob-like structures is typically solved with a Difference-of-Gaussian (DoG) detector which is integrated in the Scale Invariant Feature Transform (SIFT) (Lowe, 1999; Lowe, 2004), or a Determinant-of-Hessian (DoH) detector which is the basis for deriving Speeded-Up Robust Features (SURF) (Bay et al., 2006; Bay et al., 2008). Distinctive image regions are for instance detected with a Maximally Stable Extremal Region (MSER) detector (Matas et al., 2002). Accounting for non-incremental changes between images with similar content and thus possibly significant changes in scale, the use of a scale-space representation as introduced for the SIFT and SURF detectors is inevitable. While the SIFT and SURF detectors rely on a Gaussian scale-space, the use of a nonlinear scale-space has been proposed for detecting KAZE features (Alcantarilla et al., 2012) or Accelerated KAZE (A-KAZE) features (Alcantarilla et al., 2013).

For feature description, the main idea consists of deriving keypoint descriptors that allow to discriminate the extracted keypoints very well. Being inspired by investigations on biological vision, the descriptor presented as second part of the Scale Invariant Feature Transform (SIFT) (Lowe, 1999; Lowe, 2004) is one of the first and still one of the most powerful feature descriptors. Since, for applications focusing on computational efficiency, the main limitation of deriving SIFT descriptors consists of the computational effort, a more efficient descriptor has been presented with the Speeded-Up Robust Features (SURF) descriptor (Bay et al., 2006; Bay et al., 2008). In contrast to these descriptors consisting of a vector representation encapsulating floating numbers, a significant speed-up is typically achieved by involving binary descriptors such as the Binary Robust Independent Elementary Feature (BRIEF) descriptor (Calonder et al., 2010).

For many applications, it is important to derive stable keypoints and keypoint descriptors which are invariant to image scaling and image rotation, and robust w.r.t. image noise, changes in illumination and small changes in viewpoint. Satisfying these constraints, SIFT features are commonly applied in a variety of applications which becomes visible in more than 9.2k citations of (Lowe, 1999) and more than 29.5k citations of (Lowe, 2004), while the use of SURF features has been reported in more than 5.4k citations of (Bay et al., 2006) and more than 5.9k citations of (Bay et al., 2008).[1] Both SIFT and SURF features are also typically used for detecting feature correspondences between intensity images derived for terrestrial laser scans. For each feature correspondence, the respective keypoints may be projected to 3D space by considering the respective range information. This, in turn, yields sparse point sets of corresponding 3D points.

2.2 Keypoint-Based Point Cloud Registration

Once sparse point sets of corresponding 3D points have been derived for two scans, the straightforward solution consists of estimating a rigid-body transformation in the least squares sense (Arun et al., 1987; Umeyama, 1991). However, the two 3D point sets may also contain some point pairs resulting from incorrect feature correspondences and, consequently, it is advisable to involve the well-known RANSAC algorithm (Fischler and Bolles, 1981) for obtaining an increased robustness (Barnea and Filin, 2007; Boehm and Becker, 2007).

In case of two coarsely aligned 3D point sets, the well-known Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992) and its variants (Rusinkiewicz and Levoy, 2001; Gressin et al., 2013) may be applied. The main idea of such an approach is to iteratively minimize a cost function representing the difference between the respective sparse 3D point sets. However, if the coarse alignment between the considered 3D point sets is not good enough, the ICP algorithm may fail to converge or even get stuck in a local minimum instead of the global one. Consequently, such an approach is mainly applied for fine registration.

A further alternative consists of exploiting the spatial information of the derived 3D point sets for a geometric constraint matching based on the 4-Points Congruent Sets (4PCS) algorithm (Aiger et al., 2008) which has recently been presented with the Keypoint-based 4-Points Congruent Sets (K-4PCS) algorithm (Theiler et al., 2013; Theiler et al., 2014). While the K-4PCS algorithm provides a coarse alignment which is good enough to proceed with an

[1] These numbers were assessed via Google Scholar on 30 April 2015.

ICP-based fine registration, the processing time for the geometric constraint matching significantly increases with the number of points in the 3D point sets due to an evaluation of best matching candidates among point quadruples of both 3D point sets.

A different strategy has been presented by transferring the task of point cloud registration to the task of solving the Perspective-n-Point (PnP) problem which, in turn, may be achieved by introducing a virtual camera and backprojecting the sparse 3D point sets onto its image plane (Weinmann et al., 2011; Weinmann and Jutzi, 2011). Thus, 3D/2D feature correspondences are derived and provided as input for an efficient RANSAC-based scheme solving the PnP problem. Being robust due to accounting for both 3D and 2D cues, and being efficient due to involving a non-iterative method with only linear complexity, such an approach is still among the most accurate and most efficient approaches for registering sparse 3D point sets.

3 METHODOLOGY

As shown in Figure 1, our framework for automatically aligning TLS point clouds consists of two major steps: (*i*) feature extraction and matching and (*ii*) point cloud registration. The respective methods involved in these components are provided as well and described in the following subsections.

3.1 Keypoint Extraction and Matching

Generally, the performance of keypoint matching is an interplay of the applied keypoint detector and descriptor (Dahl et al., 2011). Hence, different keypoint detector-descriptor combinations may be applied and these may differ in their suitability, depending on the requirements of the respective application. Focusing on scan representations in the form of panoramic intensity images, where keypoint descriptors have to cope with significant changes in rotation and scale for changes in the scanner position, we only involve scale and rotation-invariant keypoint representations as listed in Table 1. More details on the respective keypoint detectors and descriptors are provided in the following.

Variant	Detector	Detector type	Descriptor	Descriptor size / type / information
1	SIFT	Blobs	SIFT	128 / float / gradient
2	SURF	Blobs	SURF	64 / float / gradient
3	ORB	Corners	ORB	32 / binary / intensity
4	A-KAZE	Blobs	M-SURF	64 / float / gradient
5	SURF*	Blobs	BinBoost	32 / binary / gradient

Table 1. The keypoint detector-descriptor combinations involved in our framework.

3.1.1 SIFT: For keypoint detection, the Scale Invariant Feature Transform (SIFT) (Lowe, 1999; Lowe, 2004) relies on convolving the image $\mathcal{I}$ and subsampled versions of $\mathcal{I}$ with Gaussian kernels of variable scale in order to derive the Gaussian scale-space. Subtracting neighboring images in the Gaussian scale-space results in the Difference-of-Gaussian (DoG) pyramid, where extrema in a $(3 \times 3 \times 3)$ neighborhood correspond to keypoint candidates. These keypoint candidates are improved by an interpolation based on a 3D quadratic function in the scale-space in order to obtain subpixel accurate locations in image space. Furthermore, those keypoint candidates with low contrast which are sensitive to noise as well as those keypoint candidates located along edges which can hardly be distinguished from each other are discarded.

In the next step, each keypoint is assigned its dominant orientation which results for the respective scale by considering the local gradient orientations weighted by the respective magnitude as well as a Gaussian centered at the keypoint. Subsequently, the local gradient information is rotated according to the dominant orientation in order to achieve a rotation invariant keypoint descriptor. The descriptor itself is derived by splitting the local neighborhood into 4×4 subregions. For each of these subregions, an orientation histogram with 8 angular bins is derived by accumulating the gradient orientations weighted by the respective magnitude as well as a Gaussian centered at the keypoint. The concatenation of all histogram bins and a subsequent normalization yield the final 128-dimensional SIFT descriptor. For deriving feature correspondences, SIFT descriptors are typically compared by considering the ratio of Euclidean distances of a SIFT descriptor belonging to a keypoint in one image to the nearest and second nearest SIFT descriptors in the other image. This ratio indicates the degree of similarity and thus the distinctiveness of matched features.

3.1.2 SURF: Speeded-Up Robust Features (SURF) (Bay et al., 2006; Bay et al., 2008) are based on a scale-space representation of the Hessian matrix which is approximated with box filters, so that the elements of the Hessian matrix may efficiently be evaluated at a very low computational cost using integral images. Thus, distinctive features in an image correspond to locations in the scale-space where the determinant of the approximated Hessian matrix reaches a maximum in a $(3 \times 3 \times 3)$ neighborhood. The detected maxima are then interpolated in order to obtain subpixel accurate locations in image space.

Similar to SIFT, a dominant orientation is calculated for each keypoint. For this purpose, the Haar wavelet responses in x- and y-direction within a circular neighborhood are weighted by a Gaussian centered at the keypoint and represented in a new 2D coordinate frame. Accumulating all responses within a sliding orientation window covering 60° yields a local orientation vector, and the orientation vector of maximum length indicates the dominant orientation. For obtaining a rotation invariant keypoint descriptor, the local gradient information is rotated according to the dominant orientation. Then, the local neighborhood is divided into 4×4 subregions and, for each subregion, the Haar wavelet responses in x'- and y'-direction are weighted by a Gaussian centered at the keypoint. The concatenation of the sum of Haar wavelet responses in x'- and y'-direction as well as the sum of absolute values of the Haar wavelet responses in x'- and y'-direction for all subregions and a subsequent normalization yield the final 64-dimensional SURF descriptor. The comparison of SURF descriptors is the same as for SIFT descriptors.

3.1.3 ORB: The approach presented with the Oriented FAST and Rotated BRIEF (ORB) detector and descriptor (Rublee et al., 2011) represents a combination of a modified FAST detector and a modified BRIEF descriptor.

The Features from Accelerated Segment Test (FAST) detector (Rosten and Drummond, 2005) analyzes each pixel (x, y) of an image $\mathcal{I}$ and takes into account those pixels located on a surrounding Bresenham circle. The intensity values corresponding to those pixels on the surrounding Bresenham circle are compared to the intensity value $\mathcal{I}(x, y)$. Introducing a threshold t, the investigated pixel (x, y) represents a candidate keypoint if a certain number of contiguous pixels have intensity values above $\mathcal{I}(x, y) + t$ or below $\mathcal{I}(x, y) - t$. A subsequent non-maximum suppression avoids keypoints at adjacent pixels. The modification resulting in the ORB detector is based on employing a scale pyramid of the image, producing FAST features at each level in the pyramid and adding an orientation component to the standard FAST detector.

The Binary Robust Independent Elementary Feature (BRIEF) descriptor (Calonder et al., 2010) is derived by computing binary

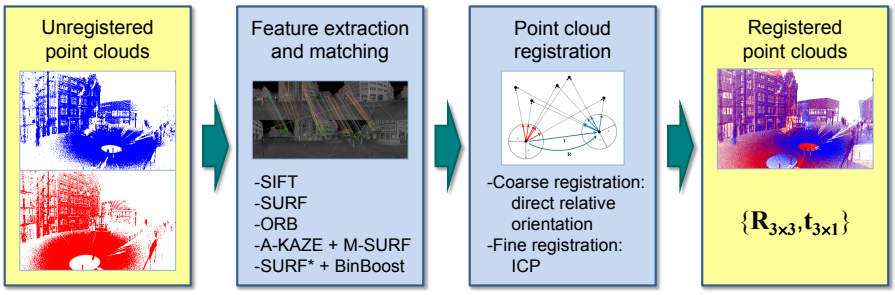

Figure 1. The proposed framework for keypoint-based point cloud registration and the involved methods for each component.

strings from image patches. In this context, the individual bits are obtained from a set of binary tests based on comparing the intensities of pairs of points along specific lines. The modification resulting in the ORB descriptor consists of steering BRIEF according to the orientation of keypoints and thus deriving a rotation-aware version of the standard BRIEF descriptor. The similarity of such binary descriptors can be evaluated by using the Hamming distance, which is very efficient to compute.

3.1.4 A-KAZE and M-SURF: Instead of the Gaussian scale-space of an image, using a non-linear scale-space may be favorable as Gaussian blurring does not respect the natural boundaries of objects and smoothes to the same degree both details and noise, reducing localization accuracy and distinctiveness of features (Alcantarilla et al., 2012). Such a non-linear scale-space may for instance be derived by using efficient additive operator splitting (AOS) techniques and variable conductance diffusion which have been employed for detecting KAZE features (Alcantarilla et al., 2012). The nonlinear diffusion filtering, in turn, makes blurring locally adaptive to the image data and thus reduces noise while retaining object boundaries. However, AOS schemes require solving a large system of linear equations to obtain a solution. In order to increase computational efficiency, it has been proposed to build a nonlinear scale-space with fast explicit diffusion (FED) and thereby embed FED schemes in a pyramidal framework with a fine-to-coarse strategy. Using such a non-linear scale-space, Accelerated KAZE (A-KAZE) features (Alcantarilla et al., 2013) may be extracted by finding maxima of the scale-normalized determinant of the Hessian matrix, where the first and second order derivatives are approximated by means of Scharr filters, through the nonlinear scale-space. After a subsequent non-maximum suppression, the remaining keypoint candidates are further refined to subpixel accuracy by fitting a quadratic function to the determinant of the Hessian response in a (3×3) image neighborhood and finding its maximum.

In the next step, scale and rotation invariant feature descriptors may be derived by estimating the dominant orientation of the keypoint in analogy to the SURF descriptor and rotating the local image neighborhood accordingly. Based on the rotated neighborhood, using the Modified-SURF (M-SURF) descriptor (Agrawal et al., 2008) adapted to the non-linear scale-space has been proposed which, compared to the original SURF descriptor, introduces further improvements due to a better handling of descriptor boundary effects and due to a more robust and intelligent two-stage Gaussian weighting scheme (Alcantarilla et al., 2012).

3.1.5 SURF* and BinBoost: Finally, we also involve a keypoint detector-descriptor combination which consists of applying a modified variant of the SURF detector and using the BinBoost descriptor (Trzcinski et al., 2012; Trzcinski et al., 2013). In comparison to the standard SURF detector (Bay et al., 2006; Bay et al., 2008), the modified SURF detector, denoted as SURF* in our paper, iterates the parameters of the SURF detector until a desired number of features is obtained[2]. Once appropriate parameters have been derived, the dominant orientation for each keypoint is calculated and used to rotate the local image neighborhood in order to allow a scale and rotation invariant feature description.

As descriptor, the BinBoost descriptor (Trzcinski et al., 2012; Trzcinski et al., 2013) is used which represents a learned low-dimensional, but highly distinctive binary descriptor, where each dimension (each byte) of the descriptor is computed with a binary hash function that was sequentially learned using Boosting. The weights as well as the spatial pooling configurations of each hash function are learned from training sets consisting of positive and negative gradient maps of image patches. In general, Boosting combines a number of weak learners in order to obtain a single strong classifier. In the context of the BinBoost descriptors, the weak learners are represented by gradient-based image features that are directly applied to intensity image patches. During the learning stage of each hash function, the Hamming distance between image patches is optimized, i.e. it is decreased for positive and increased for negative patches. Since, in our experiments, the BinBoost descriptor with 32 bytes worked best, we only report the results for this descriptor version.

3.2 Point Cloud Registration

Introducing a superscript ^{j} which indicates the respective scan $\mathcal{S}_j$ and a subscript $_i$ which indicates the respective feature correspondence, the forward-projection of n corresponding 2D keypoints $\mathbf{x}_i^1 \leftrightarrow \mathbf{x}_i^2$ between the panoramic intensity images of two scans $\mathcal{S}_1$ and $\mathcal{S}_2$ according to the respective range information yields sparse point sets of corresponding 3D points $\mathbf{X}_i^1 \leftrightarrow \mathbf{X}_i^2$. Classically, the task of keypoint-based point cloud registration is solved by estimating the rigid Euclidean transformation between the two sets of corresponding 3D points, i.e. a rigid-body transformation of the form

$$\mathbf{X}_i^2 \approx \hat{\mathbf{X}}_i^2 = \mathbf{R}_1^2 \mathbf{X}_i^1 + \mathbf{t}_1^2 \tag{1}$$

with a rotation matrix $\mathbf{R}_1^2 \in \mathbb{R}^{3\times3}$ and a translation vector $\mathbf{t}_1^2 \in \mathbb{R}^3$ (where the superscript indicates the target coordinate frame and the subscript indicates the current coordinate frame). Accordingly, the rigid-body transformation is estimated in object space.

As omnidirectional representations in the form of panoramic range and intensity images are available, we propose to estimate the transformation in observation space, i.e. we intend to find the relative orientation between consecutive scans directly. For this purpose, we apply a spherical normalization $\mathsf{N}(\cdot)$ which normalizes 3D points $\mathbf{X}_i^j$ given in the local coordinate frame of scan $\mathcal{S}_j$ to unit length and thus yields the so-called bearing vectors

$$\mathbf{v}_i^j = \mathsf{N}(\mathbf{X}_i^j) = \frac{\mathbf{X}_i^j}{\|\mathbf{X}_i^j\|} \tag{2}$$

[2]This modified version is part of OpenCV 2.4.

that simply represent the direction of a 3D point $\mathbf{X}_i^j$ w.r.t. the local coordinate frame of the laser scanner. Thus, the task of point cloud registration may be transferred to the task of finding the transformation of one set of bearing vectors to another. In photogrammetry and computer vision, this is known as relative orientation and the transformation is encoded both in the essential matrix $\mathbf{E}$ and the fundamental matrix $\mathbf{F}$, respectively. The relationship between both is given by:

$$\mathbf{F} = \mathbf{K}^{-T}\mathbf{E}\mathbf{K}^{-1} = \mathbf{K}^{-T}[\mathbf{t}]_\times \mathbf{R}\mathbf{K}^{-1} \tag{3}$$

where $[\mathbf{t}]_\times$ denotes the skew symmetric matrix of the translation and $\mathbf{K}$ represents a calibration matrix. For omnidirectional or panoramic images, the standard fundamental matrix $\mathbf{F}$ cannot be estimated, since it encapsulates the calibration matrix $\mathbf{K}$ which, in turn, is based on perspective constraints that do not hold in the omnidirectional case. The essential matrix $\mathbf{E}$, however, is independent of the camera type and may hence be used to estimate the transformation between two panoramic images.

In general, at least five points are necessary to calculate a finite number of solutions for $\mathbf{E}$ (Philip, 1998). Various algorithms for the estimation of the essential matrix $\mathbf{E}$ exist, ranging from the minimal five-point algorithms (Nistér, 2004; Stewénius et al., 2006) to the six-point (Pizarro et al., 2003), the seven-point (Hartley and Zisserman, 2008), and the eight-point (Longuet-Higgins, 1987) algorithms. As observed in seminal work (Stewénius et al., 2006; Rodehorst et al., 2008) and verified by own experiments, the five-point solver of (Stewénius et al., 2006) performs best in terms of numerical precision, special motions or scene characteristics and resilience to measurement noise. For this reason, we focus on the use of this algorithm in our paper.

3.2.1 Coarse registration: The input to our registration procedure is represented by putative feature correspondences between 2D points $\mathbf{x}_i^j$ that are subsequently transformed to bearing vectors $\mathbf{v}_i^j$ by exploiting the respective range information. Since the 2D points $\mathbf{x}_i^j$ are localized with subpixel accuracy, a respective bilinear interpolation is applied on the 2D scan grid in order to obtain the respective 3D coordinates $\mathbf{X}_i^j$. Then, the essential matrix $\mathbf{E}$ is estimated using Stewénius' five-point algorithm (Stewénius et al., 2006) and thereby involving the RANSAC algorithm (Fischler and Bolles, 1981) for increased robustness. For this, we use the implementation of OpenGV (Kneip and Furgale, 2014). A subsequent decomposition of $\mathbf{E}$ yields the rotation matrix $\mathbf{R}_1^2$ and the translation vector $\hat{\mathbf{t}}_1^2$ (Hartley and Zisserman, 2008). Since the essential matrix $\mathbf{E}$ only has five degrees of freedom, the translation vector $\hat{\mathbf{t}}_1^2$ is only known up to a scalar factor s, which is indicated by a $\hat{}$ symbol. In order to recover the scale factor s, the following is calculated over all inliers:

$$s_{\text{median}} = \underset{i}{\text{median}} \left(\|\mathbf{X}_i^2 - \mathbf{R}_1^2 \mathbf{X}_i^1\| \right) \tag{4}$$

The median is used to diminish potential outliers that could still reside in the data. Finally, the direction vector $\hat{\mathbf{t}}_1^2$ is scaled by s_{median} to get the final translation $\mathbf{t}_1^2 = s_{\text{median}} \hat{\mathbf{t}}_1^2$.

3.2.2 Fine registration: In order to remove those 3D points indicating potential outlier correspondences from the 3D point sets, we apply a simple heuristic. First, the point set $\mathbf{X}_i^1$ is transformed to $\hat{\mathbf{X}}_i^2$ using Equation 1 and the coarse estimates for $\mathbf{R}_1^2$ and $\mathbf{t}_1^2$. Then, the Euclidean distance between all corresponding 3D points is calculated and only those points with an Euclidean distance below 1m are kept. To remove such heuristics, one could employ iterative reweighted least squares techniques or a RANSAC-based modification of the ICP algorithm.

The remaining 3D points of the sparse point sets are provided to a standard ICP algorithm (Besl and McKay, 1992) which generally converges to the nearest local minimum of a mean square distance metric, where the rate of convergence is high for the first few iterations. Given an appropriate coarse registration delivering the required initial values for $\mathbf{R}_1^2$ and $\mathbf{t}_1^2$, even a global minimization may be expected. In our experiments, we apply an ICP-based fine registration and consider the result after 10 iterations.

4 EXPERIMENTAL RESULTS

In our experiments, we use a standard benchmark TLS dataset (Section 4.1) and focus on the performance of different methods for each component of the framework (Section 4.2). Additionally, we discuss the derived results w.r.t. pros and cons of the involved methods (Section 4.3).

4.1 Dataset

The involved TLS dataset[3] has been captured with a Riegl LMS-Z360i laser scanner in an area called "Holzmarkt" which is located in the historic district of Hannover, Germany. According to (Brenner et al., 2008), the Riegl LMS-Z360i has a single shot measurement accuracy of 12mm and its field-of-view covers $360^\circ \times 90^\circ$, while the measurement range reaches up to 200m. Furthermore, the angular resolution is about 0.12° and, thus, a full scan results in $3000 \times 750 = 2.25$M scanned 3D points.

In total, the dataset consists of 20 scans of which 12 were taken with (approximately) upright scan head and 8 with a tilted scan head. The single scan positions for the upright scans have been selected systematically along a trajectory with a spacing of approximately 5m, whereas the scan positions for the tilted scans almost coincide with the scan position for an upright scan, and reference values for both position and orientation have been obtained by placing artificial markers in the form of retro-reflective cylinders in the scene and carrying out a manual alignment based on these artificial targets. Thus, errors in the range of a few millimeters may be expected. In our experiments, we consider the similarity between upright and tilted scans acquired at almost the same position as too high to allow a fair statement on the registration accuracy obtained with our framework (since the respective errors w.r.t. the estimated scan position are significantly below the measurement accuracy of 12mm), and hence we only use the upright scans (Figure 2).

Since both range and intensity information are recorded for each point on the discrete scan raster, we may easily characterize each scan with a respective panoramic range image and a respective panoramic intensity image, where each image has a size of 3000×750 pixels. As the captured intensity information depends on the device, we adapt it via histogram normalization to the interval $[0, 255]$ in order to obtain 8-bit gray-valued images.

4.2 Experiments

Our experiments focus on the successive pairwise registration of scan pairs $\mathcal{P}_j = \{\mathcal{S}_j, \mathcal{S}_{j+1}\}$ with $j = 1, \ldots, 11$. For this purpose, we apply the different methods for feature extraction as described in Section 3.1 and the registration scheme as described in Section 3.2. We use the implementations provided in OpenCV 2.4 for SIFT, SURF and ORB, while we use the implementations provided with the respective paper for the other two keypoint detector-descriptor combinations. An example showing feature correspondences derived via the combination of an A-KAZE detector and an M-SURF descriptor is provided in Figure 3.

[3]This dataset and others have been released at http://www.ikg.uni-hannover.de/index.php?id=413&L=de (accessed: 30 April 2015)

Figure 2. Map of the Hannover "Holzmarkt": the position of buildings is visualized in dark gray and the scan positions for different scans $\mathcal{S}_j$ are indicated with red spots. The scan IDs are adapted according to (Brenner et al., 2008).

For evaluating the performance of our framework, the respective position and angle errors after coarse and fine registration are visualized in Figure 4 and Figure 5. Thereby, the position error indicates the deviation of the estimated scan position from the reference values, whereas the angle error has been determined by transforming the estimated rotation matrix and the respective reference to Rodrigues vectors which, in turn, allow to derive angle errors as the difference of these Rodrigues vectors w.r.t. their length (Kneip and Furgale, 2014). Furthermore, we provide the number of correspondences used for coarse and fine registration in Figure 6 in order to quantify differences between the different methods for feature extraction and matching. For coarse registration, we further provide the ratio of inliers w.r.t. all feature correspondences as well as the number of RANSAC iterations in Figure 7. In order to obtain an impression on the computational effort on a standard notebook (Intel Core i7-3630QM, 2.4Ghz, 16GB RAM), the average processing times for different subtasks are provided in Table 2 as well as the expected time for the whole process of aligning two scans. Finally, we also provide a visualization of registered TLS scans in Figure 8.

Method	t_{FEX} [s]	t_{FM} [s]	t_{CR} [s]	t_{FR} [s]	t_{Σ} [s]
SIFT	3.055	10.479	0.084	0.012	16.684
SURF	0.644	0.414	0.081	0.022	1.805
ORB	0.138	2.023	0.015	0.021	2.336
A-KAZE + M-SURF	2.815	2.533	0.007	0.050	8.221
SURF* + BinBoost	2.423	0.067	0.025	0.013	4.950

Table 2. Average processing times t_{FEX} for feature extraction, t_{FM} for feature matching, t_{CR} for coarse registration, t_{FR} for fine registration and average total time t_{Σ} required for automatically aligning two scans.

4.3 Discussion

The results provided in Figure 4 and Figure 5 reveal that the position errors after fine registration are less than 0.06m for almost all keypoint detector-descriptor combinations when considering the scan pairs $\mathcal{P}_1, \ldots, \mathcal{P}_{10}$, where the distance between the respective scan positions is between 4m and 6m. Since the respective angle errors after fine registration are below 0.15° with only a few exceptions, we may conclude that the presented method for coarse registration represents a competitive method in order to coarsely align the given scans, since a respective outlier rejection based on 3D distances is sufficient for an ICP-based fine registration. The applicability of our method for coarse registration is even further motivated by the fact that the respective processing

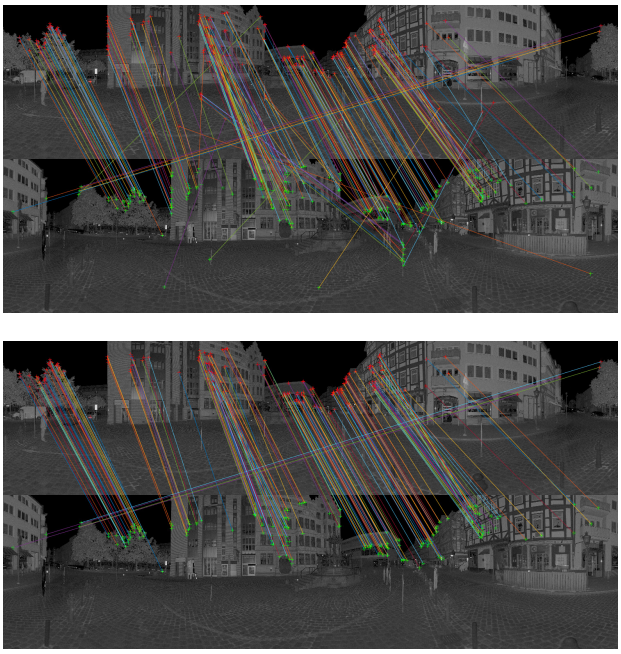

Figure 3. Feature correspondences between the panoramic intensity images of scans $\mathcal{S}_1$ and $\mathcal{S}_2$ when using the combination of an A-KAZE detector and an M-SURF descriptor: all correspondences (top) vs. inlier correspondences (bottom).

time is less than 0.085s for the considered scan pairs (Table 2). Thus, we reach a total time of less than 10s for the registration of the considered scan pairs for four of the five tested keypoint detector-descriptor combinations, and only the involved implementation for SIFT is not that efficient. Note that the time for feature extraction is counted twice, since this task is required for both scans of a scan pair.

Considering the five involved keypoint detector-descriptor combinations, we may state that A-KAZE + M-SURF and SURF* + BinBoost tend to provide the best results after fine registration (Figure 4 and Figure 5). Note that only these combinations are also able to derive a suitable position and angle estimate for the last scan pair $\mathcal{P}_{11} = \{\mathcal{S}_{11}, \mathcal{S}_{12}\}$, where the distance between the respective scan positions is approximately 12m. The respective position errors for A-KAZE + M-SURF and SURF* + BinBoost after fine registration are 0.054m and 0.085m, while the angle errors are 0.056° and 0.083°, respectively. In contrast, SIFT, SURF and ORB provide a position error of more than 0.20m and an angle error of more than 0.75° for that case.

In Figure 6, it becomes visible that the number of feature correspondences used for coarse registration is similar for SIFT, ORB and SURF* + BinBoost, while it tends to be higher for SURF. For A-KAZE + M-SURF, even a significant increase of this number may be observed across all 12 scan pairs. The increase in the number of involved feature correspondences for A-KAZE + M-SURF compared to the other keypoint detector-descriptor combinations is even more significant when considering fine registration, where it is partially even more than twice as much as for the others. Based on these characteristics, an interesting trend becomes visible when considering the respective ratio of inliers during coarse registration. While the inlier ratio is comparable for SIFT and SURF, it is better for SURF* + BinBoost, and it is considerably better for ORB and A-KAZE + M-SURF (Figure 7, top). A high percentage of inliers, in turn, has a positive impact on coarse registration by significantly reducing the number of RANSAC iterations (Figure 7, bottom). Consequently, the

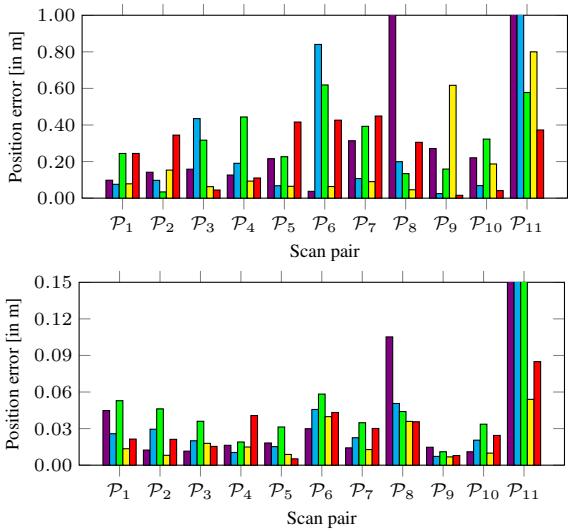

Figure 4. Position errors after the coarse registration (top) and after the fine registration (bottom) of scan pairs $\mathcal{P}_j = \{\mathcal{S}_j, \mathcal{S}_{j+1}\}$: SIFT (violet), SURF (cyan), ORB (green), A-KAZE + M-SURF (yellow), SURF* + BinBoost (red).

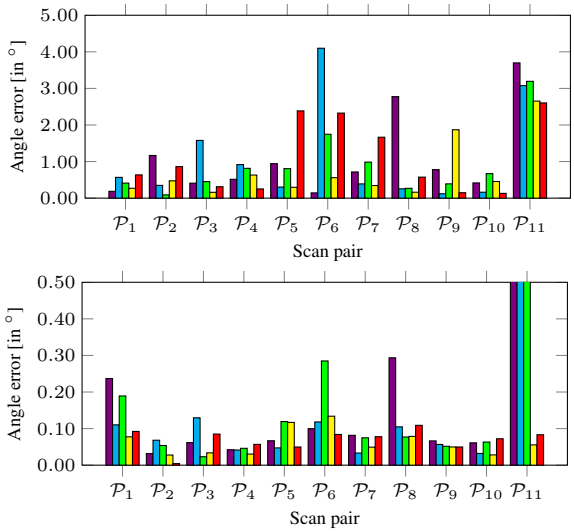

Figure 5. Average angle errors after the coarse registration (top) and after the fine registration (bottom) of scan pairs $\mathcal{P}_j = \{\mathcal{S}_j, \mathcal{S}_{j+1}\}$: SIFT (violet), SURF (cyan), ORB (green), A-KAZE + M-SURF (yellow), SURF* + BinBoost (red).

combination A-KAZE + M-SURF not only increases the number of correspondences, but also the inlier ratio and, thus, the position and angle errors across all scan pairs tend to be the lowest for this combination (Figure 4 and Figure 5). For most of the scan pairs, the respective position error is even close or below the given measurement accuracy of 12mm.

Finally, we may state that our framework is suited for both urban environments and scenes containing vegetation, and it does neither depend on regular surfaces nor human interaction. The only limitation may be identified in the fact that feature correspondences have to be derived between the panoramic intensity images derived for the respective scans. In this regard, we may generally observe that the total number of feature correspondences decreases with an increasing distance between the respective scan positions and, accordingly, the quality of the registration results will decrease. However, this constraint holds for the other image-based approaches as well and is not specific for our framework.

5 CONCLUSIONS

In this paper, we have presented a novel framework for evaluating the influence of different keypoint detector-descriptor combinations on the results of point cloud registration. While we involve five different approaches for extracting local features from the panoramic intensity images derived for the single scans, the registration process has been transferred from object space to observation space by considering the forward-projection of putative feature correspondences and exploiting bearing vectors instead of the corresponding 3D points themselves. Our results clearly reveal that replacing SIFT and SURF detectors and descriptors by more recent approaches significantly alleviates point cloud registration in terms of accuracy, efficiency and robustness.

For future work, we plan to integrate more approaches for feature extraction as well as more approaches for keypoint-based point cloud registration in our framework in order to objectively evaluate their performance on publicly available benchmark TLS datasets. In this context, it would also be desirable to point out chances and limitations of the different approaches w.r.t. different criteria specified by potential end-users, e.g. the spacing between adjacent scans or the complexity of the observed scene. Promising results may be expected.

REFERENCES

Agrawal, M., Konolige, K. and Blas, M. R., 2008. CenSurE: center surround extremas for realtime feature detection and matching. *Proceedings of the European Conference on Computer Vision*, Vol. IV, pp. 102–115.

Aiger, D., Mitra, N. J. and Cohen-Or, D., 2008. 4-points congruent sets for robust pairwise surface registration. *ACM Transactions on Graphics* 27(3), pp. 1–10.

Alcantarilla, P. F., Bartoli, A. and Davison, A. J., 2012. KAZE features. *Proceedings of the European Conference on Computer Vision*, Vol. VI, pp. 214–227.

Alcantarilla, P. F., Nuevo, J. and Bartoli, A., 2013. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *Proceedings of the British Machine Vision Conference*, pp. 13.1–13.11.

Arun, K. S., Huang, T. S. and Blostein, S. D., 1987. Least-squares fitting of two 3-D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 9(5), pp. 698–700.

Barnea, S. and Filin, S., 2007. Registration of terrestrial laser scans via image based features. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVI-3/W52, pp. 32–37.

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

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

Bay, H., Tuytelaars, T. and Van Gool, L., 2006. SURF: speeded up robust features. *Proceedings of the European Conference on Computer Vision*, Vol. 1, pp. 404–417.

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

Boehm, J. and Becker, S., 2007. Automatic marker-free registration of terrestrial laser scans using reflectance features. *Optical 3-D Measurement Techniques* VIII, pp. 338–344.

Brenner, C., Dold, C. and Ripperda, N., 2008. Coarse orientation of terrestrial laser scans in urban environments. *ISPRS Journal of Photogrammetry and Remote Sensing* 63(1), pp. 4–18.

Calonder, M., Lepetit, V., Strecha, C. and Fua, P., 2010. BRIEF: binary robust independent elementary features. *Proceedings of the European Conference on Computer Vision*, Vol. IV, pp. 778–792.

Dahl, A. L., Aanæs, H. and Pedersen, K. S., 2011. Finding the best feature detector-descriptor combination. *Proceedings of the International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pp. 318–325.

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

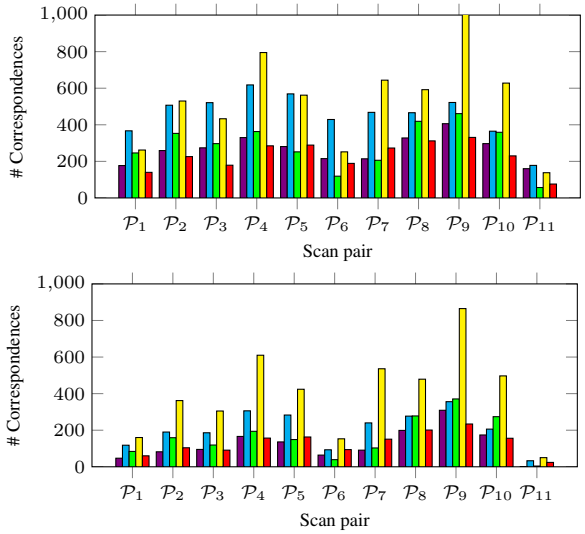

Figure 6. Number of feature correspondences used for the coarse registration (top) and for the fine registration (bottom) of scan pairs $\mathcal{P}_j = \{\mathcal{S}_j, \mathcal{S}_{j+1}\}$: SIFT (violet), SURF (cyan), ORB (green), A-KAZE + M-SURF (yellow), SURF* + BinBoost (red).

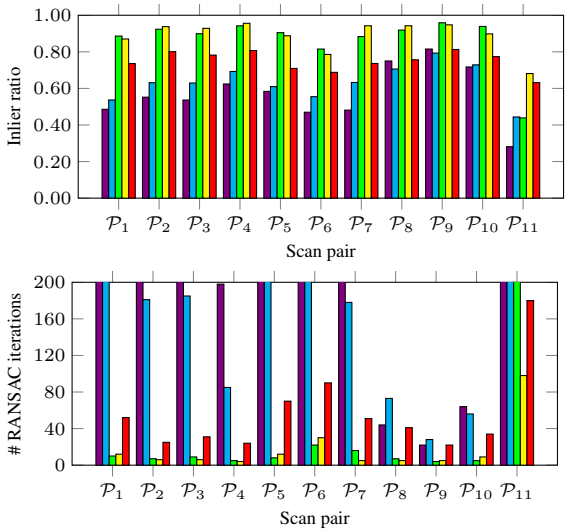

Figure 7. Inlier ratio (top) and number of RANSAC iterations (bottom) for the coarse registration of scan pairs $\mathcal{P}_j = \{\mathcal{S}_j, \mathcal{S}_{j+1}\}$: SIFT (violet), SURF (cyan), ORB (green), A-KAZE + M-SURF (yellow), SURF* + BinBoost (red).

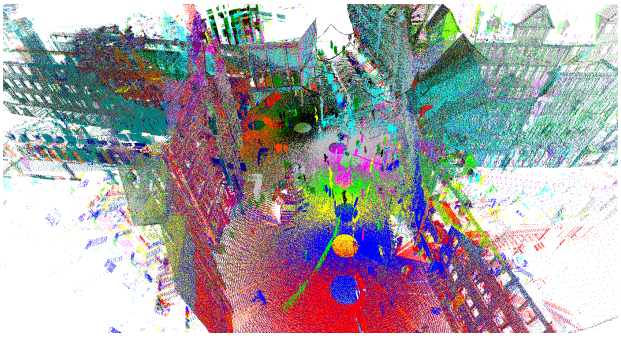

Figure 8. Aligned point clouds when using A-KAZE + M-SURF: the points belonging to different scans $\mathcal{S}_j$ are encoded with different color.

Gressin, A., Mallet, C., Demantké, J. and David, N., 2013. Towards 3D lidar point cloud registration improvement using optimal neighborhood knowledge. *ISPRS Journal of Photogrammetry and Remote Sensing* 79, pp. 240–251.

Harris, C. and Stephens, M., 1988. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151.

Hartley, R. I. and Zisserman, A., 2008. *Multiple view geometry in computer vision*. University Press, Cambridge, UK.

Kneip, L. and Furgale, P., 2014. OpenGV: a unified and generalized approach to real-time calibrated geometric vision. *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1–8.

Longuet-Higgins, H. C., 1987. A computer algorithm for reconstructing a scene from two projections. In: Fischler, M. A. and Firschein, O. (Eds.), *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*. Morgan Kaufmann, San Francisco, USA, pp. 61–62.

Lowe, D. G., 1999. Object recognition from local scale-invariant features. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1150–1157.

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

Matas, J., Chum, O., Urban, M. and Pajdla, T., 2002. Robust wide baseline stereo from maximally stable extremal regions. *Proceedings of the British Machine Vision Conference*, pp. 36.1–36.10.

Nistér, D., 2004. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6), pp. 756–770.

Philip, J., 1998. *Critical point configurations of the 5-, 6-, 7-, and 8-point algorithms for relative orientation*. Technical Report TRITA-MAT-1998-MA-13, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden.

Pizarro, O., Eustice, R. and Singh, H., 2003. Relative pose estimation for instrumented, calibrated imaging platforms. *Proceedings of Digital Image Computing Techniques and Applications*, pp. 601–612.

Rodehorst, V., Heinrichs, M. and Hellwich, O., 2008. Evaluation of relative pose estimation methods for multi-camera setups. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVII-B3b, pp. 135–140.

Rosten, E. and Drummond, T., 2005. Fusing points and lines for high performance tracking. *Proceedings of the International Conference on Computer Vision*, Vol. 2, pp. 1508–1515.

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., 2011. ORB: an efficient alternative to SIFT or SURF. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2564–2571.

Rusinkiewicz, S. and Levoy, M., 2001. Efficient variants of the ICP algorithm. *Proceedings of the International Conference on 3D Digital Imaging and Modeling*, pp. 145–152.

Stewénius, H., Engels, C. and Nistér, D., 2006. Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing* 60(4), pp. 284–294.

Theiler, P. W., Wegner, J. D. and Schindler, K., 2013. Markerless point cloud registration with keypoint-based 4-points congruent sets. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. II-5/W2, pp. 283–288.

Theiler, P. W., Wegner, J. D. and Schindler, K., 2014. Keypoint-based 4-points congruent sets – Automated marker-less registration of laser scans. *ISPRS Journal of Photogrammetry and Remote Sensing* 96, pp. 149–163.

Trzcinski, T., Christoudias, M., Fua, P. and Lepetit, V., 2012. Learning image descriptors with the boosting-trick. *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vol. 1, pp. 269–277.

Trzcinski, T., Christoudias, M., Fua, P. and Lepetit, V., 2013. Boosting binary keypoint descriptors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2874–2881.

Tuytelaars, T. and Mikolajczyk, K., 2008. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision* 3(3), pp. 177–280.

Umeyama, S., 1991. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(4), pp. 376–380.

Weinmann, M., 2013. Visual features – From early concepts to modern computer vision. In: Farinella, G. M., Battiato, S. and Cipolla, R. (Eds.), *Advanced Topics in Computer Vision*. Advances in Computer Vision and Pattern Recognition, Springer, London, UK, pp. 1–34.

Weinmann, M. and Jutzi, B., 2011. Fully automatic image-based registration of unorganized TLS data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVIII-5/W12, pp. 55–60.

Weinmann, M., Weinmann, M., Hinz, S. and Jutzi, B., 2011. Fast and automatic image-based registration of TLS data. *ISPRS Journal of Photogrammetry and Remote Sensing* 66(6), pp. S62–S70.