Accurate Single Image Multi-Modal Camera Pose Estimation

Christoph Bodensteiner, Marcus Hebel, and Michael Arens

Fraunhofer IOSB,
Gutleuthausstrae 1, 76275 Ettlingen, Germany
{christoph.bodensteiner,marcus.hebel,michael.arens}@iosb.fraunhofer.de
http://www.iosb.fraunhofer.de

Abstract. A well known problem in photogrammetry and computer vision is the precise and robust determination of camera poses with respect to a given 3D model. In this work we propose a novel multi-modal method for single image camera pose estimation with respect to 3D models with intensity information (e.g., LiDAR data with reflectance information).
We utilize a direct point based rendering approach to generate synthetic 2D views from 3D datasets in order to bridge the dimensionality gap. The proposed method then establishes 2D/2D point and local region correspondences based on a novel self-similarity distance measure. Correct correspondences are robustly identified by searching for small regions with a similar geometric relationship of local self-similarities using a Generalized Hough Transform. After backprojection of the generated features into 3D a standard Perspective-n-Points problem is solved to yield an initial camera pose. The pose is then accurately refined using an intensity based 2D/3D registration approach.
An evaluation on Vis/IR 2D and airborne and terrestrial 3D datasets shows that the proposed method is applicable to a wide range of different sensor types. In addition, the approach outperforms standard global multi-modal 2D/3D registration approaches based on Mutual Information with respect to robustness and speed.
Potential applications are widespread and include for instance multi-spectral texturing of 3D models, SLAM applications, sensor data fusion and multi-spectral camera calibration and super-resolution applications.

Key words: Multi-Modal Registration, Pose Estimation, Multi-Modal 2D/3D Correspondences, Self-Similarity Distance Measure

1 Introduction

A fundamental issue in computer vision and photogrammetry is the precise determination of camera poses with respect to a given 3D model. It has many applications, e.g., augmented reality, image based localization or robot navigation. The involved registration task is mostly formulated as the determination of a geometric transformation[1] which maps corresponding features onto each other

[1] In case of camera pose estimation the geometric transformation is known as the external calibration matrix or extrinsic parameters of the camera

by minimizing a proper distance measure. In general there are two solution approaches for matching 2D/3D image data. Either one computes 3D information from 2 or more 2D images and performs the similarity comparison in 3D, or 2D data is simulated from the 3D dataset and compared in a two-dimensional space. We focus on the latter, since we assume only one available 2D image and a prerecorded 3D dataset with intensity information as described in Sec.2. This paper considers 2D/3D camera pose estimation for multi-modal data, i.e., estimating the external camera $R, \mathbf{t}$ parameters when the internal camera parameters K are known and the involved datasets stem from different image modalities. The projection of 3D world points $\mathbf{M_i}$ to corresponding 2D image points $\mathbf{m_i}$ is modeled by a standard pinhole camera model. The intrinsic parameters K with the parameters skew s, focal length f, aspect ratio α and principal point $\mathbf{u} = [u_0\ v_0]^T$ are assumed to be known.

$$\mathbf{m_i} = P\mathbf{M_i}, P = K\left[R|\mathbf{t}\right], K = \begin{bmatrix} f & s & u_0 \\ 0 & \alpha f & v_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1}$$

1.1 Related Work and Contribution

2D/3D camera pose estimation received much attention in the last decades [1–4]. Existing methods can be roughly divided by the spatial extent/type of the used features/structures:

Pose from 2D/3D Point Correspondences: Pose estimation is basically solvable from 3 2D/3D point correspondences and is widely known as the $P3P$ problem. A common approach is to determine the 3D point positions $\mathbf{M_i^C}$ in the camera coordinate frame C. This leads to a root finding problem for a polynomial of degree 8 with only even terms. To disambiguate the 4 solutions in the general case an additional point is often used. However, the computed pose from 4 point correspondences is usually not accurate and therefore it is advisable to simultaneously use $n >> 4$ point correspondences. This leads to the well known PnP (perspective n points) problem [3]. Often RANSAC type algorithms [5] or robust cost functions [3] are used to handle outliers in the correspondence set. A non-linear least squares optimization of the reprojection error with all inlying feature correspondences increases accuracy further:

$$\text{minimize}_{R,\mathbf{t}} \sum_i \|K(R\mathbf{M_i} + \mathbf{t}) - \mathbf{m_i}\|_2^2 . \tag{2}$$

Modern algorithms [1, 4] efficiently solve this problem under real-time constraints even on modest computing hardware [4].

Pose from Planar Structures: By observing a corresponding planar structure in both datasets one can extract the pose parameters directly from the homography H [6, 7] which maps the structures onto each other. In this case the

projection equation of model $\mathbf{M_i}$ and 2D image points $\mathbf{m_i}$ simplifies without loss of generality to:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} \mathbf{r_1} & \mathbf{r_2} & \mathbf{r_3} & \mathbf{t} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = K \begin{bmatrix} \mathbf{r_1} & \mathbf{r_2} & \mathbf{t} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \tag{3}$$

where $\mathbf{r_i}$ denotes the $i.th$ column of the matrix R. Therefore the model points $\mathbf{M_i}$ and image points $\mathbf{m_i}$ are related by a homography H (defined up to a scale factor λ):

$$H = [\mathbf{h_1} \ \mathbf{h_2} \ \mathbf{h_3}] = \lambda K [\mathbf{r_1} \ \mathbf{r_2} \ \mathbf{t}]. \tag{4}$$

Based on the assumption that K is known, the camera pose is given by:

$$\begin{aligned} \mathbf{q_1} &= \lambda K^{-1} \mathbf{h_1} \\ \mathbf{q_2} &= \lambda K^{-1} \mathbf{h_2} \\ \mathbf{q_3} &= \mathbf{r_1} \times \mathbf{r_2} \\ \mathbf{t} &= \lambda K^{-1} \mathbf{h_3} \end{aligned} \tag{5}$$

Due to data noise the computed matrix $Q = [\mathbf{q_1} \ \mathbf{q_2} \ \mathbf{q_3}]$ usually does not satisfy the ortho-normality constraint of a rotation matrix $R, R^T R = I$. Therefore R is computed to minimize$_R \|R - Q\|_F^2$ s.t. $R^T R = I$ in a Frobenius norm sense. This can be efficiently achieved [6] by a singular value decomposition of $Q = U \ S \ V^T$ and setting R to $U \ V^T$.

Pose from Intensity Based Distance Minimization: A standard approach for pose determination in the field of medical image computing (e.g., X-Ray/CT-computed tomography, X-Ray/MR-magnetic resonance imaging) is to simulate pose parametrized 2D views $V_{sim}(R, \mathbf{t})$ from the 3D dataset which minimize/maximize an intensity based distance/similarity measure $D_{(Typ)}, D : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ between the acquired reference image I_R and a simulated view over the support of the image region A.

$$\text{minimize}_{R,\mathbf{t}} \int_A D_{(Typ)}(V_{sim}(R, \mathbf{t}), I_R). \tag{6}$$

We refer to [8, 9] for a comparison of common intensity based 2D/3D distance measures e.g., Normalized Cross Correlation (NCC), Spearman Rank Order Correlation (SPROCC), Gradient Correlation (GC), Correlation Ratio (CR) and Mutual Information (MI).

Generally, intensity based similarity optimization allows for accurate registration results but is computationally expensive. Additionally, these methods often rely on a very good initialization to avoid local optima. Local feature methods are more advantageous when significant changes of the underlying scenery hamper global intensity based similarity computations. However, a common difficulty of the outlined approaches is the determination of 2D/3D feature/planar

region correspondences, respectively a sufficiently close starting point for an intensity based similarity computation. Local feature based correspondence methods [10, 11] work very well if the image data stems from the same image modality. An excellent review can be found in [12]. Local feature approaches mostly match common image features based on gradient information. The registration task becomes challenging if the image data is multi-modal, e.g., the image intensity data stems from different sensors with, e.g., different image acquisition techniques, spectral sensitivities or passive/active illumination. The problem of finding accurate local feature correspondences across different image modalities is less understood. Successful multi-modal matching applications mostly stem from medical image registration, e.g., the fusion of MR/CT or CT/PET (positron emission tomography) images by maximization of the information theoretic similarity measure MI.We focus on the determination of point and region correspondences using local multi-modal features. The main difficulty is the inherent trade off between feature correspondence discrimination and multi-modal matching capabilities. We adapt the approach of Shechtman and Irani [13] who proposed self-similarity descriptors for sketch based object and video detection and extend it with ideas from the work of Leibe et. al. [14] to determine multi-modal point and region correspondences. To the best of our knowledge there is no literature about accurate multi-modal pose determination with local correspondences based on self-similarity. We additionally propose to refine the pose optimization by minimizing locally a densely computed self-similarity distance to accurately align local image regions where standard multi-modal similarity measures like MI or CR have major difficulties. The fusion of 2D images with LiDAR data is still an active research field [15–17]. The closest work [18] with respect to our application uses MI to register optical images with LiDAR data. However, we claim that our method is more robust w.r.t. to pose initialization and cluttered image data.

The outline of the paper is as follows: first we give a short overview for laser based acquisition of 3D data. Then we describe the key parts of the approach and discuss specific details which enable the robust local correspondence search in the multi-modal case. We evaluate the method on different image datasets with a focus on IR/Vis in combination with airborne (ALS) and terrestrial laser scanning (TLS) datasets. In the end, we discuss the results and give further research directions.

2 Laser-based 3D Data Acquisition

Remote sensing of 3D structures in the far-field is commonly approached with multi-view image analysis as well as active illumination techniques. In this context, LiDAR (light detection and ranging) is a comparatively new method that enables direct acquisition of 3D information [19]. LiDAR sensors emit laser radiation and detect its reflection in order to determine the precise distance between sensor and illuminated object. Currently available laser scanners are capable of performing hundreds of thousands of range measurements per second, thus

allowing a complete 3D scenery to be captured in a reasonably short time interval. Two main types of laser scanners can be distinguished that follow different concepts of range determination: phase shift and time-of flight laser scanners. In case of phase-shift scanners, a continuous laser beam is emitted with sinusoidally modulated optical power. The distance to the reflecting object is estimated based on the phase shift between received and emitted signal. Phase-shift scanners are well suited for static terrestrial laser scanning. When operating the scanning head on a rigid tripod, ranging accuracies of few millimeters at distances up to hundred meters can be achieved. Mobile methods like airborne laser scanning usually combine a time-of-flight LiDAR device with high-precision navigational sensors mounted on a common sensor platform. The ranging accuracy of such a system is typically limited to few centimeters, while maximum distances up to one kilometer can be measured.

Currently available time-of-flight laser scanners are capable of acquiring the full waveform of reflected pulses, thus enabling new methods of data analysis [20]. The portion of the reflected energy can be considered in relation to the emitted radiation and the measured distance. This ratio reveals the local reflectivity at the specific laser wavelength, which typically lies in the near infrared due to eye-safety reasons. High-speed scanning and exploitation of reflectivity information results in highly detailed textured 3D point clouds. However, unlike ambient background light, the reflection of directed laser radiation is significantly affected by the incidence angle and the surface characteristics of the illuminated objects.

3 Method

The multi-modal 2D/3D registration procedure can be summarized as follows: first we utilize a point based rendering approach to generate a synthetic 2D View from the 3D dataset to enable the correspondence search. Then we establish 2D/2D point and local region correspondences based on local features. Correct correspondences are robustly identified by searching for small regions with a similar geometric relationship of local features by employing a Generalized Hough Transform. The 3D positions for the synthetically generated 2D features can easily be determined using the depth buffer information from the rendering procedure. The registration is then carried out by solving a PnP based pose determination. The calculated pose is finally refined with an intensity based registration. This refinement step is intended for applications with very high accuracy requirements, e.g., multi-spectral texturing of 3D models, multi-modal camera calibration or multi-modal super-resolution. Summarizing, the method can be divided (cf., Fig.1) as follows:

1. Synthetic 2D View Generation
2. Feature Extraction
3. Feature Correspondence Search and Constraint Filtering
4. Feature Correspondence Based Pose Determination
5. Intensity Based Multi-Modal Registration

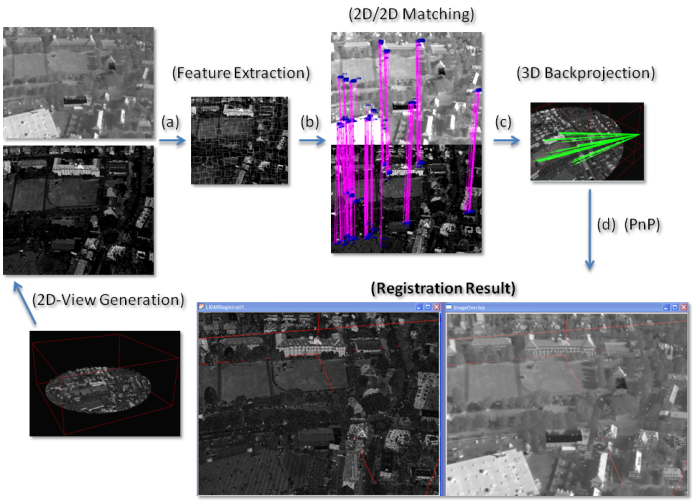


Fig. 1. Registration algorithm overview: (a) extracted local image regions, (b) feature matching by searching geometrically consistent feature matches and fundamental matrix filtering, (c) 3D backprojection of 2D feature matches, (d) pose determination. The registration result (bottom image) shows a superposition (red cross) of the airborne IR image and the textured LiDAR view from the left side.

Synthetic 2D View Generation: We propose a direct point based rendering approach [21] for synthetic view generation. The automated generation of texture mapped models (e.g., Fig. 4i) is still error prone and a time consuming process. To this end we use a simple rendering of the 3D point cloud data based on small spheres with adaptive sizes. In this work we selected the initial pose for the view generation manually. However, the proposed feature based method shows a wide convergence range.

Feature Extraction: We extract local features over different scales and use standard descriptors for an initial correspondence search. To this date we evaluated SIFT [10], SURF [11] and recently proposed self-similarity descriptors [13].

Feature Correspondence Search and Constraint Filtering: To enable a robust local feature based 2D/3D registration approach for multi-modal data we utilize the concept of simultaneously matching local features inside small image regions. The selection of these image regions serves as a starting point for the correspondence search. Each region defines a local coordinate frame, where the geometric layout of contained features is determined. Our experiments show that it is favorable to use image regions with strongly distinct features in order to increase the number of correct region matches. In this work we used

constant region sizes (60x60px) and a simple heuristic based on Harris corners, which serve as the origin. We employ a Generalized Hough Transform similar to the one used in [14]. We also use a technique called soft-matching [14] for local feature matching which incorporates the k (e.g., 2-4) nearest neighbors in descriptor space as potential matches. Due to fundamentally different ob-

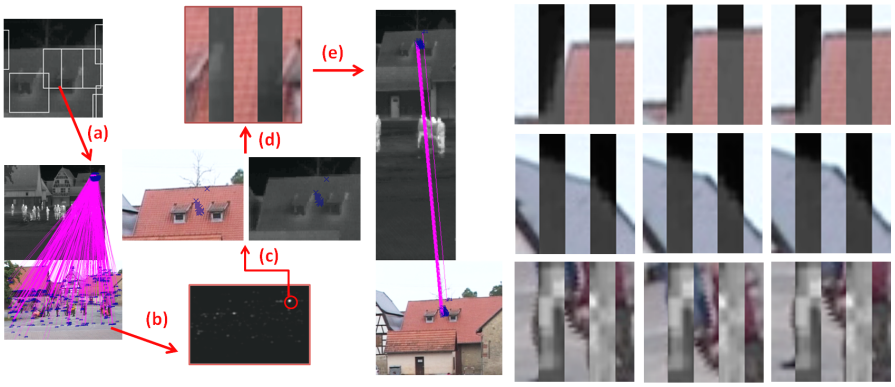


Fig. 2. Local feature correspondence search algorithm (IR/Vis example): (a) extracted local image regions (Harris/Foerstner), (b) feature matching by searching geometrically consistent feature matches, (c) best hypothesis supporting feature matches, (d) intensity based local image region alignment, (e) final point correspondences for one local image region. The right columns shows local patch alignments (Init/MI/Self-Sim).

ject appearances, many initial local feature matches are not correct and would lead to an enormous amount of wrong point correspondences (see Fig. 2a left). Therefore each local feature casts a vote for a corresponding region center according to the geometric layout in its reference coordinate system [14]. Under the assumption that wrong correspondences spread their votes randomly, we determine the corresponding image region center with a simple maximum search. The final 2D/2D point correspondences are feature matches which contributed a vote near to the maximum in voting space. We refer to Leibe et. al. [14] for a detailed description of the voting principle. However, the voting space maximum in multi-modal image pairs does not always correspond to a correct region match. We use point correspondences that contributed a vote near the maximum in voting space (backprojection of best hypothesis supporting feature matches) to estimate (RANSAC) a local affine transformation T_a of the corresponding image patch (e.g., 60x60px). We then discard matches with a high self-similarity distance (eqn. 12) based on an empirical determined threshold.

Intensity Based Optimization of Local Planar Patches: Due to small errors in the determined feature correspondences we also applied a local intensity

based multi-modal distance optimization to find local region correspondences (cf., Fig.2right). Formally, we search for a set of optimal transformation parameters $\hat{\theta}$ which minimize a multi-modal distance measure $D : \mathbb{R}^N \rightarrow \mathbb{R}$ over the support of a local image region A_i around the determined point correspondences:

$$\hat{\theta} = argmin_\theta \; D(\theta), \tag{7}$$

$$D(\theta) = \int_{A_i} D_{(\cdot)}(I_{T_\theta}, I_R). \tag{8}$$

To this end we use parametric (projective) transformations T_θ with 8 degrees of freedom for distance minimization. The local image region A_i should be as small as possible for projective transformations since they inherently imply planarity. The affine transformation T_a serves as a starting point for the image alignment optimization. The nonlinear optimization is based on a specific pattern search method which does not rely on gradient information. Basically, we approximate the distance function $D(\theta)$ with a multi-variate polynomial of degree 2 and recenter/rescale a search pattern at the optimum of the surrogate polynomial. Given a proper initialization, the method needs only a few distance function evaluations to converge to a local optimum and is especially designed for computationally expensive distance functions. We plan to directly compute an accurate pose from the local projective transformations T_θ as described in Sec. 1.1. However, to this end we use this computationlly expensive step only for an optional point correspondence optimization, when we omit a global intensity based similarity optimization.

Feature Correspondence Based Pose Determination: The corresponding 3D feature positions from the 2D rendering are efficiently backprojected into 3D by using the depth buffer information from the rendering process[2]. Given the 2D/3D correspondences we calculate the pose using a standard PnP algorithm. We used the recently proposed EPnP [4] algorithm, which expresses the n 3D feature positions as a weighted sum of four virtual control points. This algorithm proved to be superior w.r.t. speed and accuracy compared to the popular POSIT algorithm [22]. To robustly detect outliers in the 2D/3D correspondence set we employed a RANSAC approach. We used $n = 8$ subset sizes and a $5px$ reprojection error (cf. eqn. 2) threshold for the inlier set. The computed pose was additionally refined by a non-linear Gauss-Newton minimization of the reprojection error (eqn. 2) w.r.t. the inlier set.

Intensity Based Multi-Modal Registration: To accurately align the multi-modal data sets we additionally minimize/maximize an intensity based distance/similarity measure. The convergence range of intensity based multi-modal 2D/3D methods is usually very small. However, the local feature based pose

[2] It's important to transform the data into an adequate coordinate system to reduce inaccuracies caused by a limited Z-Buffer resolution

computation usually provides a sufficiently close starting point. An important design choice is the selection of an appropriate distance measure. Mutual Information [9] is considered the gold standard similarity measure for multi-modal matching. It measures the mutual dependence of the underlying image intensity distributions:

$$D_{(MI)}(I_R, I_{T_\theta}) = H(I_R) + H(I_{T_\theta}) - H(I_R, I_{T_\theta}) \tag{9}$$

where $H(I_R)$ and $H(I_{T_\theta})$ are the marginal entropies and

$$H(I_R, I_{T_\theta}) = \sum_{X \in I_{T_\theta}} \sum_{Y \in I_R} p(X,Y) log(\frac{p(X,Y)}{p(X)p(Y)}) \tag{10}$$

is the joint entropy. $p(X,Y)$ denotes the joint probability distribution function of the image intensities X, Y in I_R and I_{T_θ}, and $p(X)$ and $p(Y)$ are the marginal probability distribution functions. However, MI is very difficult to estimate (e.g.,see Fig. 3) for small image regions and does not cope well with spatially-varying intensity fluctuations (eqn. (10)). Therefore we propose to minimize a self-similarity distance of corresponding image regions in I_R and I_{T_θ}. To compute the self-similarity description for an image patch point we compare a small image patch with a larger surrounding image region centered at $q \in R_i$ using simple sum of squared differences (SSD) between image intensities normalized by the image patch intensity variance and noise $c(I)_{noise,variance}$:

$$S_q(x,y) = \exp\left(-\frac{SSD_q(x,y)}{c(I)_{noise,variance}}\right) \tag{11}$$

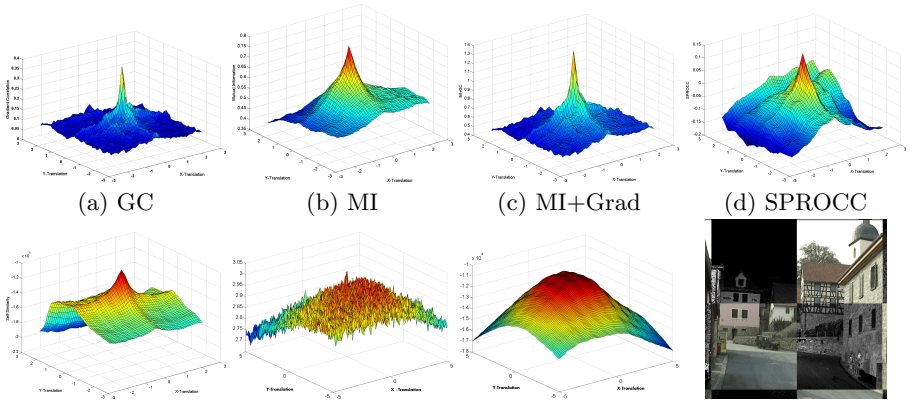
This correlation image $S_q(x,y)$ is then transformed into a log polar coordinate system and partitioned into bins (e.g., 20 angles, 4 radial intervals) where the maximal correlation value in each bin is used as an entry for the self-similarity dimension description of the vector $S_q^{I(\cdot)}(x,y)$ located at the image position $(x,y) \in A_i$. Each vector is then linearly normalized to $[0,1]$. The distance measure now simply computes the sum of squared distances of the self-similarity description vectors $S_q^{I(\cdot)}$ computed at the region A_i:

$$D_{(SSim)}(I_{T_\theta}, I_R) = \sum_{(x,y) \in A_i} \| S_q^{I_R}(x,y) - S_q^{I_{T_\theta}}(x,y) \|^2. \tag{12}$$

In multiple experiments we plotted the values of the optimization function while varying function parameters as shown in Fig. 3f,g. The plots of this distance measure show unique maxima and relatively smooth and monotonically increasing function shapes especially for small local image regions.

3.1 Implementation Details and Runtime Information

The implemented point based rendering and intensity based 2D/2D and 2D/3D registration software is based on the OpenCV and VTK [23] C/C++ libraries.



(a) GC (b) MI (c) MI+Grad (d) SPROCC

(e) Self-Sim (f) MI-local (g) Self-Sim-local (h) Optimal Pose

Fig. 3. Plots of global distance/similarity values (a-e) for deviations from the found value by the optimization algorithm (camera translation in x and y direction (±2.5m)). Comparison of local MI and Self-Sim values (f,g) for a small image patch (60x60px) (image translation in x and y direction) from the found value by the optimization algorithm. The optimal value for the pose (h) is shown at where all parameters are zero.

We used the SiftGPU [24] and OpenSURF [11] implementation for the local descriptor computation. Since the voting based correspondence approach requires many feature correspondence searches, it is important to use fast search structures [25] for nearest neighbor determination in descriptor space (L_2 norm). Since our implementation is not runtime-optimized, the reported time measurements provide only a rough estimate for the actual overall algorithm runtime. The determination of 500 region correspondences ranges from 60-200s on an Intel Q9550 System. An intensity based local image patch refinement (60x60px) needs 10-15s (single core) for one image patch optimization (Self-Similarity features and distance measure).

4 Results

We evaluated the voting based feature correspondence method (Sec.3) by counting point correspondences w.r.t. a robustly estimated fundamental matrix (8-pt algorithm, RANSAC, 1.25px inlier threshold). When possible, e.g., in case of IR/Vis aerial images we estimated a global homography (RANSAC, 2.5px inlier threshold) to evaluate correct correspondences. In total we used 10 Vis/IR, 50 Vis/IR aerial and 2 LiDAR/IR/Vis image pairs. Our experiments show (Tab. 1, Fig. 4) that this method enables a robust determination of multi-modal feature correspondences. The self-similarity descriptors proved to be well suited for this task compared to well established local feature approaches like SIFT [10] or SURF [11] (see Fig. 4). A visualization comparing SIFT, SURF and

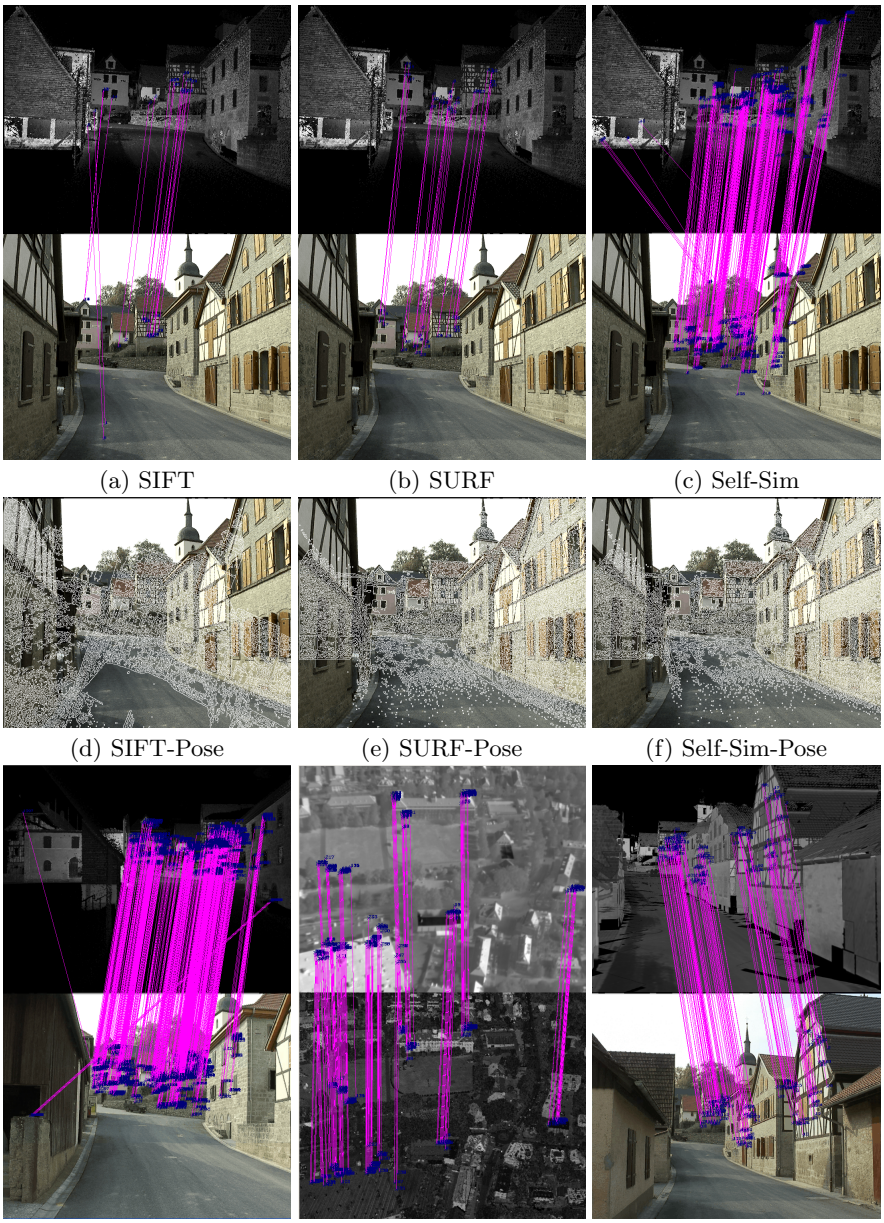
Features	IR/Vis (2D/2D)	IR/Vis (Aerial 2D/2D)	ALS/IR
SIFT	0 / 0	63 / 85%	0 / 0
SURF	0 / 0	35 / 91%	0 / 0
Self-Sim	3706 / 49%	2185 / 64%	4881 / 23%

Table 1. Averaged rounded (found/correct) point correspondences. The correctness of point correspondences was additionally checked by visual inspection in case of fundamental matrix constraint filtering.

Self-Similarity features for TLS/Vis image data is shown in Fig.4. This effect especially holds for ALS based renderings from close view points where rendering holes drastically affect gradient histogram based descriptors (e.g., Fig.4h). Given high quality synthetic renderings and high point densities local feature methods based on gradient information can still work. Fig.4(a-f) shows correspondences and PnP based pose computations for SIFT, SURF and Self-Sim features. However, the number and distribution of correct correspondences was considerably higher for Self-Sim features. In case of IR/ALS (cf., Fig.4h) data we were not able to compute correct correspondences using standard local features like SIFT and SURF. To evaluate the pose determination accuracy from the found point correspondences we calculated ground truth pose information by jointly matching small sets of 3-5 images in order to calculate accurate extrinsic and intrinsic parameters. Then we artificially perturbed the camera positions from T_{World}^{Cam} to $T_{World}^{pertCam}$. The translation parameters were randomly perturbed by maximally ± 5 m and the rotation parameters were perturbed by maximally ± 3 deg. After registration we calculated the Euler angle representation of the deviation matrix T_{dev} using the calculated registration matrix T_{regCam}^{World}.

$$T_{dev} = T_{regCam}^{World} T_{World}^{Cam}. \tag{13}$$

The average point based pose estimation accuracy for 20 TLS/Vis views showed rotational deviations of 0.95 (x), 1.12 (y) and 0.74 (z) degree and an average translational deviation of 0.93 (x), 0.72 (y), 0.69 (z) m for voting based SURF feature correspondences. Self-Similarity feature correspondences led to rotational deviations of 0.89 (x), 1.02 (y) and 0.97 (z) degree and an average translational deviation of 0.89 (x), 0.93 (y), 0.67 (z) m. For the intensity based multi-modal registration, we evaluated various intensity based distance measures like MI, CR, GC, SPROCC, linear combinations of MI+GC and the proposed densely computed Self-Similarity. First we evaluated intensity based registration performance for local patches by visual inspection with respect to MI, Spearman Rank Correlation Coefficient and Self-Similarity. We evaluated the number of (correct/false) alignments for a representative set of 113 local image patches. SPROCC led to 34% correct alignments, MI to 31%, and Self-Sim to 94% correct alignments (e.g., Fig.2right). By using the global intensity based 2D/3D camera pose estimation step we finally achieved very accurate visual registration results (cf., Fig. 5).

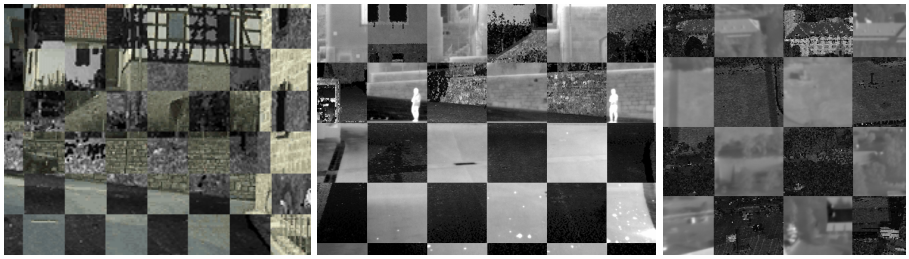


(a) SIFT (b) SURF (c) Self-Sim

(d) SIFT-Pose (e) SURF-Pose (f) Self-Sim-Pose

(g) TLS/Vis (h) IR/ALS (i) TLS (TextureMap)/Vis

Fig. 4. Voting based correspondences using (a) SIFT, (b) SURF and (c) Self-Similarity features for identical Vis/TLS images. The second row (d-f) shows resulting poses using the PnP approach. The last row depicts Self-Similarity feature correspondences for TLS/Vis (g), IR/ALS (h) and TLS(Texture Mapped)/Vis (i) data. All correspondences are fundamental matrix constraint (RANSAC, 2.25px) filtered.



(a) Vis/TLS Result (b) IR/TLS Result (c) IR/ALS Result

Fig. 5. Intensity based camera pose estimation results for Vis/TLS-LiDAR (a), IR/TLS-LiDAR (b) and IR/ALS-LiDAR (c) image pairs.

5 Conclusion and Future Work

In this work we proposed and implemented a robust method to determine accurate local multi-modal 2D/3D correspondences. The method is based on simultaneously matching geometrically consistent feature correspondences. Very accurate multi-modal 2D/3D alignments can be achieved in combination with local intensity based optimization which allows for a precise multi-spectral texturing of 3D models, sensor data fusion and multi-spectral camera calibration.

The registration of multi-modal 2D/3D datasets is inherently difficult due to fundamental differing object appearances. Multi-modal distance measures are usually application dependent and the suitability of self-similarity as a general multi-modal distance measure remains open. However, experiments show a clear dominance of the proposed self-similarity distance measure for IR/Vis and ALS/TLS/IR/Vis image pairs in case of small region sizes (see Fig.2right). In addition, we find the approach of locally matching self-similar structures [13] very intriguing since it does not assume a global functional relationship like correlation ratio or clusters in the joint intensity distribution like MI [9]. Most importantly Self-Sim copes well with spatially varying intensity fluctuations. Future research directions are manifold. The fast computation of self-similarity descriptors and distances is crucial for the practicability of the method. Moreover, we work on an extension of the voting procedure to enable wide baseline scenarios. We also plan to extend the method to allow a robust and accurate multi-modal 2D/3D registration starting from a sparsely sampled set of 2D renderings of large scale 3D models without any knowledge of extrinsic and intrinsic camera parameters.

References

1. Lu, C.P., Hager, G.D., Mjolsness, E.: Fast and globally convergent pose estimation from video images. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 610–622

2. David, P., DeMenthon, D., Duraiswami, R., Samet, H.: Softposit: Simultaneous pose and correspondence determination. International Journal of Computer Vision **59** (2004) 259–284
3. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press, ISBN: 0521540518 (2004)
4. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate o(n) solution to the pnp problem. International Journal of Computer Vision **81** (2009) 155–166
5. Raguram, R., Frahm, J.M., Pollefeys, M.: A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In: ECCV. (2008)
6. Zhang, Z.: A flexible new technique for camera calibration. Technical report, Microsoft Research (1998)
7. Benhimane, S., Malis, E.: Homography-based 2d visual tracking and servoing. The International Journal of Robotics Research **26** (2007) 661–667
8. Penney, G., Weese, J., Little, J.A., Desmedt, P., Hill, D.L., Hawkes, D.J.: A comparison of similarity measures for use in 2-d-3-d medical image registration. IEEE Transactions on Medical Imaging **17** (1998) 586–595
9. Viola, P., Wells, W.: Alignment by maximization of mutual information. International Journal of Computer Vision **24** (1997) 137–154
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
11. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding **110** (2008) 346–359
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1615–1630
13. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: CVPR. (2007)
14. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. International Journal of Computer Vision **77** (2008) 259–289
15. Vasile, A., Waugh, F.R., Greisokh, D., Heinrichs, R.M.: Automatic alignment of color imagery onto 3d laser radar data. In: AIPR. (2006)
16. Ding, M., Lyngbaek, K., Zakhor, A.: Automatic registration of aerial imagery with untextured 3d lidar models. In: CVPR. (2008)
17. Wang, L., Neumann, U.: A robust approach for automatic registration of aerial images with untextured aerial lidar data. In: CVPR. (2009)
18. Mastin, A., Kepner, J., Fisher, J.: Automatic registration of lidar and optical images of urban scenes. In: CVPR. (2009)
19. Vosselman, G., Maas, H.G.: Airborne and Terrestrial Laser Scanning. Whittles Publishing, Dunbeath, Scotland (2010)
20. Wagner, W., Ullrich, A., Ducic, V., Melzer, T., Studnicka, N.: Gaussian decomposition and calibration of a novel small-footprint full-waveform digitising airborne laser scanner. ISPRS Journal of Photogrammetry and Remote Sensing **60** (2006)
21. Gross, M., Pfister, H.: Point-Based Graphics. MORGAN KAUFMANN (2007)
22. DeMenthon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. International Journal of Computer Vision **15** (1995) 123–141
23. Schroeder, W., Martin, K., Lorensen, B.: The Visualization Toolkit: An Object-Oriented Approach to 3-D Graphics. Kitware (2003)
24. Wu, C.: SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). Technical report, University of North Carolina at Chapel Hill (2007)
25. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP. (2009)