Automatic Registration of RGB-D Scans via Salient Directions

Bernhard Zeisl
ETH Zurich
zeislb@inf.ethz.ch

Kevin Köser *
GEOMAR, Kiel
kkoeser@geomar.de

Marc Pollefeys
ETH Zurich
pomarc@inf.ethz.ch

Abstract

We address the problem of wide-baseline registration of RGB-D data, such as photo-textured laser scans without any artificial targets or prediction on the relative motion. Our approach allows to fully automatically register scans taken in GPS-denied environments such as urban canyon, industrial facilities or even indoors. We build upon image features which are plenty, localized well and much more discriminative than geometry features; however, they suffer from viewpoint distortions and request for normalization.

We utilize the principle of salient directions *present in the geometry and propose to extract (several) directions from the distribution of surface normals or other cues such as observable symmetries. Compared to previous work we pose no requirements on the scanned scene (like containing large textured planes) and can handle arbitrary surface shapes. Rendering the whole scene from these repeatable directions using an orthographic camera generates textures which are identical up to 2D similarity transformations. This ambiguity is naturally handled by 2D features and allows to find stable correspondences among scans. For geometric pose estimation from tentative matches we propose a fast and robust 2 point sample consensus scheme integrating an early rejection phase. We evaluate our approach on different challenging real world scenes.*

1. Introduction

When surveying construction sites, historical buildings or industrial facilities laser scanning is the state-of-the-art technique to obtain accurate three-dimensional models. To obtain a full 3D model, several 2.5D scans have to be combined, i.e. registered to each other. Usually a scanner is positioned at different places, in- or outdoors, in order to minimize scan shadows and to obtain a model as complete as possible. Since scanning is a time-consuming and therefore expensive task the number of scans is usually kept as small as possible, leading to a wide baseline setting between the scan positions. Not only scanning, but also the registration of individual scans takes a lot of time - either afterwards by manually aligning models, or on site by carefully positioning targets (artificial markers) in the scene, which are spotted and automatically detected from several scan positions. If one desires to rescan the facility at another point in time and align current data with an older model, exploiting artificial markers for registration is impossible. As a result there is a quest for automatic registration methods which do not rely on any artificial landmarks, but can generate accurate registration results by exploiting the scan data itself.

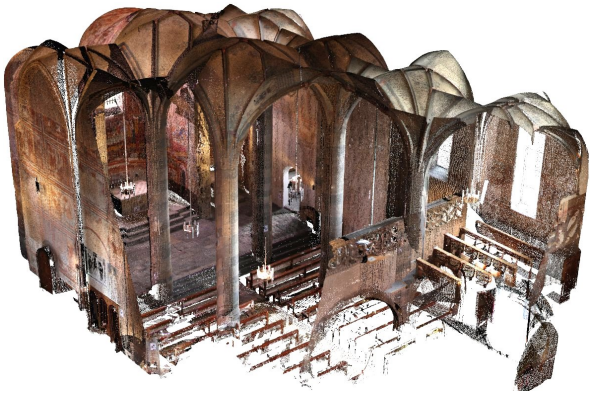


Figure 1: Cut through a 3D models obtained by our algorithm from 5 individual scans (CHURCH dataset). We achieve entirely automatic registration of arbitrary geometry from largely different viewpoints by exploiting depth and image data jointly.

Along this line local alignment methods such as ICP [2] require a good initialization and are not applicable to wide baseline scenarios or when the relative rotation is unknown. GPS and magnetic compass can simplify the registration problem, but they fail under bridges, inside buildings, urban canyon or close to metallic or electric installations, respectively. Modern laser scanners come with inbuilt or attachable cameras and deliver distance plus color information and we aim at exploiting this data jointly for fully automatic registration. For the image data the nuisance of viewpoint (position and orientation of the camera) needs to be factored

*This work was done while K. Köser was employed at the Institute for Visual Computing at ETH Zurich.

out. This is an interesting problem in its own right, important also for nowadays' consumer depth cameras (such as Kinect) or stereo systems, coined RGB-D matching in the computer vision literature [11, 22, 25].

In this work we propose to become independent of the original sensor viewpoint by exploiting characteristic *salient directions* of the scene, which are repeatable among different scans. Examples include peaks in the distribution of the surface normals, vanishing points, symmetry, gravity or other directions that can be reliably obtained from the sensor or the scene. Each salient direction is then exploited to render an *orthographic view*, and by this way removing the perspective effects that had been introduced by the particular scanner position. Importantly, for corresponding salient directions between scans generated images are identical (for jointly seen Lambertian scene parts) up to a 2D similarity transformation! Thus, standard feature detection and description approaches can be employed and features are computed in a viewpoint normalized image representation. Compared to earlier approaches proposed for consumer depth cameras [25] or stereo systems [22, 4] our approach does not pose any requirements on the presence of particular geometric shapes. Moreover, we do not rely on features only on particular fitted models (planes, cylinders, cones), but match the whole visible scene, this way significantly increasing the surface area where features can be extracted. This is an important aspect if the visible overlap between scans is small. Contrary to previous work where depth discontinuities can not be handled, our rectification approach generates images that consistently capture objects and features across different levels of depth. Such features at geometry boundaries and folds are among the most discriminative, as known e.g. from stereo. Finally, we propose a novel 2-point solution for the restricted 4 DoF registration problem, allowing for a greedy rejection of outlier-contaminated hypotheses in a sample consensus framework.

The remainder of the paper is structured as follows: After a discussion of existing registration techniques in the next section, we show how to obtain viewpoint invariance from salient directions in Sec. 3. Then Sec. 4 and Sec. 5 cover details of our approach for salient direction detection and pose estimation. Finally, we present results on real data and evaluate the new technique in Sec. 6.

2. Related Work

In this work we assume that the calibration of the capture system (camera and scanner) is given and for each scan range and image data share the same center of projection. This registration can be performed by targets or calibration patterns [21] or by maximizing mutual information between reflectance and color [17]. For registering the system's pose at two largely different positions with different orientations, related work can be classified into three categories:

Approaches utilizing image information only First, purely image based approaches build upon features which are approximately invariant against perspective distortion, such as affine features [16], or - to a lesser degree - SIFT [15] and variants thereof. Given established feature correspondences an initial 3D rigid transformation can be estimated, which can be used to bootstrap ICP [2] to obtain a refined registration. As has been argued e.g. in [11], when using affine normalization discriminative power is lost, i.e. one can no longer distinguish real world circles and ellipses. Finally, because of the strong requirements for the local region, considerably less features can be found reliably as compared to simpler detectors. In addition the affine detector is taking substantially more time.

Approaches using geometry information Second, approaches using geometry descriptors have been shown to work on 3D scenes [9, 19, 24]. A key difficulty is the estimation of the position, scale and orientation in 3D space where to compute the descriptor, i.e. a good 3D feature detector. Several detectors have been proposed, e.g. [23, 10, 8], however a major dilemma in 2.5D (as opposed to real 3D) is as follows: A useful point for matching requires a repeatable detection. Consequently surface parts need to be seen also from another - widely different - viewpoint. However, this repeatability is likely to decrease with increasing surface complexity because of self-occlusions. On the other hand for low complexity surfaces the exact localization is sensitive to noise and the local geometry is not discriminative for matching. An alternative to 3D feature detectors is to densely sample the surface, leading to a very high number of descriptors (e.g.[9]) that need to be handled in matching and verification.

Approaches building upon both modalities Finally, and in the direction of our work, there are approaches that normalize images with respect to the geometry before image feature detection and matching. For planar scenes like facades with clearly visible straight lines, vanishing points can be used, even if no depth information is available [18, 3, 1]. For more general scenes, it was shown that the sole usage of affine features can be improved, if they are normalized with respect to the local surface normal rather than to the affine shape [11]. Still this approach relies on the affine detector and shares its drawbacks. For large planar structures in a scene, viewpoint invariant patches [22, 4] can be detected and rotated to a frontal view. Recently, this local approach has been generalized from planes to parametric developable surfaces, allowing also to use cylinders and cones [25]. For complex scenes the detection of these parametric objects becomes the bottleneck of the approach. Further, the main problem still remains and matches can only be obtained on isolated objects and interesting texture has to lie on the detected geometry.

Figure 2: (left:) Images taken from two different positions, which naturally exhibit a wide baseline. The red altar visualizes correspondence. Feature matching and thus registration from these images fails in most cases. (middle, right:) Generated salient direction rectified (SDR) renderings along corresponding salient directions. Images of the roof are equivalent up to a 2D euclidean transformation (cf. Claim 2), while the right most images correspond up to a translation (cf. Claim 3).

3. Viewpoint Invariance via Salient Directions

Our novel approach to register widely separated scans builds upon image features rather than 3D geometry features, because image features are plenty, well localized and discriminative. We eliminate effects of viewpoint to allow for wide baseline registration of scans without a prediction on relative pose. In contrast to related work, we do not require the presence of particular geometric shapes (e.g. planes). Instead we exploit the entire scene information by the concept of *salient directions*.

Let us now define what we mean by a salient direction. The pose of a laser scanner in the world coordinate system is specified by the mapping of a point X from world to scanner coordinates via $X_i = s_i R_i X + t_i = s_i R_i X - s_i R_i C_i$. Here, C_i represents the origin of the scanner in world coordinates, while R_i represents its orientation and s_i is the scaling. In the following we will use the index i to refer to any single scan and indices i, j to distinguish between any two scans.

Definition 1. *A salient direction is a real-world direction in global coordinates d^{sal} that can be observed locally as $d_i^{\text{sal}}, d_j^{\text{sal}}$ in independent scans i and j:*

$$d^{\text{sal}} = R_i^T d_i^{\text{sal}} = R_j^T d_j^{\text{sal}}. \tag{1}$$

Intuitively, imagine d^{sal} is the north direction, that is represented in scans i and j as d_i^{sal} and d_j^{sal} respectively.

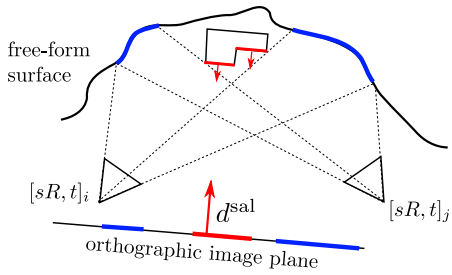


Figure 3: Orthographic renderings along a salient direction. The scene overlap of planar (red) and free form (blue) surface will be rendered identically along d^{sal} for each scanner.

As input to our algorithm we consider 2.5D depth and image data, either from a laser scanner or from a consumer depth device or stereo system. In case of panoramic data we assume that both image and depth data are given as faces of a cube-map. Then, for the depth data, local normals are estimated and we will call the set of range data, color data and normals taken from one position a *scan*. The goal is now to render a view which is suitable for matching it against other scans. Ideally we want to produce a normalized image that looks the same as a normalized image from another location (see Fig. 2 for examples and Fig. 3 for an illustration).

Definition 2. *A salient direction rectified (SDR) image, is an image which is obtained by rendering the scene along a salient direction d_i^{sal} with orthographic projection matrix*

$$P_i = \begin{bmatrix} \tilde{r}_{i,1}^T \\ \tilde{r}_{i,2}^T \end{bmatrix} \quad \text{with} \quad P_i\, d_i^{\text{sal}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{2}$$

where $\{\tilde{r}_{i,1}, \tilde{r}_{i,2}, d_i^{\text{sal}}\}$ forms an orthonormal basis of $\mathbb{R}^3$ and relates to the orthographic camera coordinate system.

Claim 1. *Given a salient direction d^{sal} with corresponding local directions $d_i^{\text{sal}}, d_j^{\text{sal}}$ in scans i and j, then corresponding points in the two SDR-images relate to each other via a 2D similarity transformation.*

As simple proof we want to show that with the given projection matrices $P_i^{\text{sal}}, P_j^{\text{sal}}$ image points x_i, x_j relate to each other via

$$x_j = s' R' x_i + t', \tag{3}$$

where s', R' and t' denote 2D scaling, rotation and translation respectively. Without loss of generality we set the i^{th} scanner pose $[I, 0]$ and denote $[s_j R_j,\, t_j] = [sR,\, t]$. Then according to Eq. (2) for a 3D point X its projections in the two SDR-images are $x_i = P_i X$ and $x_j = P_j(sRX + t)$. Also according to Eq. (2) P_i and $P_j R$ span the same basis. Thus comparison with Eq. (3) reveals

$$t' = P_j t \quad \text{and} \quad R' = P_j R P_i^{\mathrm{T}} \quad \text{and} \quad s' = s. \tag{4}$$

Since Eq. (4) holds for every point X the solution is unique. Further it is easily verified that $R'^{\mathrm{T}}R' = I$ and thus R' is orthogonal. As a result images must be related by a similarity transform, which proves the claim. SIFT features are well suited for handling this remaining ambiguity.

Claim 2. *If absolute scale is known – as for laser scans – the freedom reduces to a 2D euclidean transformation.*

The proof is trivial, since for constant scale across scenes $s = 1$. As a result this allows for scale variant feature description and matching. Observe that there is still one degree of freedom in choosing P_i, i.e. there is an undetermined in-plane rotation.

Claim 3. *Given that a global direction g is known commonly among scans in local coordinates as g_i and that $\tilde{r}_{i,1}$ is chosen as $\tilde{r}_{i,1} = (g_i \times d_i^{\mathrm{sal}})/|g_i \times d_i^{\mathrm{sal}}|$, then generated images differ only in translation.*

Defining $\tilde{r}_{i,1}$ as above and setting $\tilde{r}_{i,2}$ orthogonal to it via $\tilde{r}_{i,2} = (d_i^{\mathrm{sal}} \times \tilde{r}_{i,1})/|d_i^{\mathrm{sal}} \times \tilde{r}_{i,1}|$ ensures that g appears upright in the SDR-images. In this case $R' = I$ which leaves only t' and proves the claim. Only in case g coincides with d^{sal}, $\tilde{r}_{i,1}$ is undefined (a case which is easily spotted) and in-plane rotation is still ambiguous. In all other cases simple upright feature descriptors can be employed, which have been shown to be more discriminative than features with locally-adaptive orientation [1].

Our approach is separated into four stages, which we will explain in more detail in the next Sec. 4 and Sec. 5

1. Detection of salient directions (per scan).
2. Normalization of image data with respect to salient directions (per direction per scan)
3. Extraction of features (per SDR-image) and establishment of tentative correspondences
4. Geometric verification and concurrent pose estimation (for a scan pair)

4. Salient Direction Detection and Image Normalization

Given a salient world direction that can be identified in two different scans, we have shown that we can transform the image content in a way that it becomes virtually invariant with respect to the unknown pose. Depending on the scene type several possibilities exist how to identify salient directions, including vanishing points [1] in modern architecture, directions of repetitions or symmetries [12] in historical buildings or north direction from the sky or the time and the sun [14] in outdoor scenes. However, in this contribution we demonstrate the idea using salient directions derived from characteristics of geometric structures, that is peaks in the distribution of surface normals (cf. Fig. 4). For

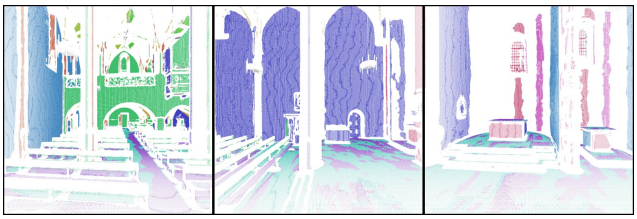


Figure 4: Support regions for different detected salient directions (color coded), shown for 3 cube faces.

successful registration only a single peak needs to be consistent, while remaining modes can be different.

Dominant normal directions Potentially disjoint, locally planar surfaces give rise to dominant surface normals. Detection of those is rephrased as finding peaks within the sampled point-normal distribution in each scan. Mean shift [5] is a suited approach to achieve this goal. It allows to model the density without explicitly parameterizing it, by evaluating a kernel K for normal n via

$$\hat{f}(n) = \frac{1}{|\mathcal{N}(n)|} \sum_{n_i \in \mathcal{N}(n)} K(n, n_i), \tag{5}$$

with $\mathcal{N}(n)$ being the set of neighbors of n. We initialize mean shift with 50 samples obtained as cluster centers from K-means. The algorithm now performs gradient descent on the density estimate $\hat{f}(n_k)$ and sample trajectories reach stable points at peaks of the density function.

As a distance measure between normals we use their orientation agreement. In particular we utilize the cosine distance $1 - n^{\mathrm{T}}n_i$ which in general relates to density estimation on a hypersphere. Furthermore, we employ a symmetric kernel with a smooth Parzen estimate (i.e. decaying weight on normals at larger distance) with an additional cut off at a maximum of $\varphi = 10°$. Thus

$$K(n, n_i) = \begin{cases} c_h \cdot \exp\left(-\frac{1}{h}(1 - n^{\mathrm{T}}n_i)\right), & n^{\mathrm{T}}n_i > \cos(\varphi) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where h is specifying the strength of the exponential weighting and c_h is a normalization constant [1].

The sampling density of points on a surface highly depends on the distance of the surface from the scanner, as well as the slant of the surface wrt. the scanning direction. Thus, if we used raw 3D points x (and their normals n) as generated from the scanner much higher emphasis would be given to surfaces close to the scanner and parallel to the scanning direction. In particular SDR-images would be ren-

[1] For two points on the unit sphere squared euclidean and cosine distance are equivalent: $\frac{1}{2}\|a - b\|^2 = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - 2a^{\mathrm{T}}b\right) = 1 - a^{\mathrm{T}}b$. Thus a also mean-shift with a symmetric Gaussian kernel with variance $\sigma^2\mathbf{I} = h\mathbf{I}$ fulfills our conditions exactly.

dered from salient directions highly supported by structures near the scanner, and repeatability of salient directions between scans would be degraded. Thus, before mode finding we re-sample the point data. Conceptually the sampling likelihood $p(x)$ of a point x is proportional to the area it describes in the 3D scene, i.e.

$$p(x) \propto a(x) \cdot \sec(\arccos\langle -r_x, n\rangle). \tag{7}$$

Here $a(x)$ denotes the surface area orthogonal to the scanning direction r_x. For a depth map it is the projected pixel footprint at depth $[x]_z$, while for a laser-scan it relates to the projected 2D scan interval (given by the angular scan resolution) at distance $\|x\|$. As a result we generate a spatially evenly sampled point cloud and are able to determine salient directions bias-free.

View Synthesis When rendering 2.5D data from a different viewpoint, missing 3D information introduces holes in the generated images. Keypoints are not detected in these visual artifacts, but descriptors might reach into or gap them. Since a descriptor captures gradient information, our desire is to avoid strong edges due to artifacts (which would perturb it) and we perform in-filing via a diffusion process. This keeps gradients small such that descriptors focus on the present texture information.

Since we don't require a fully consistent 3D mesh, we fill small holes directly in each SDR-image. Two different kinds of holes must be distinguished and in-painting is handled differently. Within the first category are holes which are caused by occluders in the original scanner viewpoint placed in front of the surface to render, e.g. a statue in front of a facade. In this case we not only fail to capture depth data (for parts of the facade in the given example) but also the corresponding texture (the texture of the statue will be captured, not the facade). As a matter of fact, in-painting is performed on the rendered SDR-images itself. Second, are holes which are caused by missing data in the scanning process (e.g. at reflective structures). Compared to the former, here texture information is available and thus we aim for a smooth inpainting on the orthographic depth-map. Then detected areas are re-rendered with the updated depth information to obtain the original texture. Both cases can be easily distinguished by back-projecting hypothesized surface points into the original views and evaluating whether or not an occluder is present. Holes themselves are detected by searching for connected components in the initially rendered images, while for inpainting we utilize FMM [20] as it is simple and fast.

5. Feature Extraction and Pose Estimation

Given several SDR-images of the scene, in each local image features are extracted. Since 3D models are given with absolute scale we can make use of Claim 2 and for each feature its absolute size is known. Thus we could even apply features of fixed size, e.g. Harris corners [7]. However, to detect (different) structures at various levels of detail we perform feature detection in scale space using DoG [15]. Still, as a constraint for matching we restrict the search for correspondences to those with same spatial extent. Similarly, in case the in-plane rotation of the orthographic camera is fixed (cf. Claim 3) we employ upright descriptors. For each feature we find tentative correspondences by fast approximate nearest neighbor search in descriptor space.

For the relative registration of two scans we augment each feature by its 3D position and normal in the local coordinate system and denote points as p_s and p_t (in the following indices s and t indicate source and target scan, respectively). Then we seek the parameters of the relative transformation $[R, t]$ from source to target. For a laser scanner the gravity direction is usually known (assumed to be aligned with the z-axis in the following), so we need to estimate only 4 parameters; however, for a hand-held RGB-D sensor 6 DoF need to be estimated. In either way, pose estimation is performed within a random sample consensus scheme, i.e. in each round the support for a generated transformation hypothesis $[R, t]$ is evaluated. Approaches presented in the following differ in the way they generate a transformation hypothesis in each iteration.

Relative Bearing and 3D Offset (4 DoF) As pointed out by Wu *et al.* [22] each point defines a local coordinate system via its normal and feature orientation. For upright features the latter is fixed by the gravity direction and the local coordinate system is defined as $[n, n \times e_z, (n \times e_z) \times n])$. Thus a single feature correspondence suffices to estimate a transformation hypothesis. The rotation angle θ around the gravity direction e_z is computed between normals n_s, n_t projected in the x-y plane

$$\begin{aligned} &\bar{n}_s = n_s - \langle n_s, e_z\rangle e_z \quad \text{and} \quad \bar{n}_t = n_t - \langle n_t, e_z\rangle e_z \\ &\theta = \arccos\langle \bar{n}_s, \bar{n}_t\rangle \cdot \operatorname{sign}\langle e_z, (\bar{n}_s \times \bar{n}_t)\rangle, \end{aligned} \tag{8}$$

while the translation is then given by $t = p_t - R_z(\theta)\, p_s$. As an alternative to RANSAC a 1D voting scheme via kernel density estimation can be employed efficiently [22].

We have found that normal vectors of extracted features tend to be noisy and are thus of limited value in their use for pose estimation. This is in particular the case for consumer depth cameras or stereo systems [2] and has two reasons. First, they are computed only in a local neighborhood and second, detected image features often correspond with structure boundaries introducing errors in the normal computation. As an alternative to using normals for registration, we will exploit the fact that corresponding local coordinate system axes can also be computed from pairs of correspon-

[2] Normals in [22] are taken from the estimated plane model, i.e. the approach fits planes rather than individual feature points.

Algorithm 1 2-point geometric pose verification

Require: set $\mathbf{m} = [m_1, \dots m_n]$, $m_i = \{p_s^{(i)}, p_t^{(i)}\}$ of M potential matches between source and target scene
Require: number of iterations K and inlier threshold ε
for $k = 1, \dots K$ **do**
  uniformly sample 2 matches m_i, m_j from $\mathbf{m}$
  $v_s \leftarrow p_s^{(i)} - p_s^{(j)}, \quad v_t \leftarrow p_t^{(i)} - p_t^{(j)}$
  if $|\|v_s\| - \|v_t\|| > \varepsilon$ **or** $|\langle v_s, e_z\rangle - \langle v_t, e_z\rangle| > \varepsilon$ **then**
    reject sample pair and **continue**
  $\bar{v}_s \leftarrow v_s - \langle v_s, e_z\rangle e_z \quad$ and $\quad \bar{v}_t \leftarrow v_t - \langle v_t, e_z\rangle e_z$
  $\theta \leftarrow \arccos\langle \bar{v}_s, \bar{v}_t\rangle \cdot \operatorname{sign}\langle e_z, (\bar{v}_s \times \bar{v}_t)\rangle$
  $t \leftarrow \frac{1}{2}\left(p_t^{(i)} - R_z(\theta)\, p_s^{(i)} + p_t^{(j)} - R_z(\theta)\, p_s^{(j)}\right)$
  for all $l \in [1, M]$ **do**
    if $\|p_t^{(l)} - R_z(\theta)\, p_s^{(l)} + t\| < \varepsilon$ **then**
      insert m_l in $\mathbf{s}$
  if $|\mathbf{s}| > |\mathbf{s}^*|$ **then**
    $\mathbf{s}^* \leftarrow \mathbf{s}, \quad [R_z^*, t^*] \leftarrow [R_z(\theta), t]$
return final transformation $[R_z^*, t^*]$ and best inlier set $\mathbf{s}^*$

dences. The orientation of these vectors is more precisely compared to normals due to their much larger spatial extent.

This gives rise to our robust 2-point geometric relative pose verification, which is presented in Alg. 1 (typical values are $K = 1000$, $\varepsilon = 3cm$). It incorporates an early rejection of generated hypotheses, such that only a fraction (on average 25% in our experiments) of generated transformation hypotheses need to be evaluated wrt. all data. Related to our new algorithm is the idea of filtering wrong correspondences in [9]; however there authors use a heuristic rather than constructing an efficient RANSAC framework.

A transformation hypothesis is formed from 2 potential matches i, j drawn at random from the correspondence set. 3D points $p_s^{(i)}$, $p_s^{(j)}$ in the source and $p_t^{(i)}$, $p_t^{(j)}$ in the target scene form vectors v_s and v_t respectively, connecting the 2 points in the local scans. If the chosen samples are correct matches, then the length of these two vectors must be equal. In addition, because we are searching for a rotation around the z-axis, their height difference has to be equal as well. This leads to an *early rejection criterion* allowing to avoid computing and testing the underlying transformation hypothesis. Given the previous two conditions hold, we first compute a relative rotation $R_z(\theta)$ from the two vectors similar to Eq. (8). Second we evaluate the translation t between target and rotated source points.

Full 6 DoF transformation To estimate all 6 DoF of a 3D rigid body transformation, at minimum 3 corresponding points are required (if normals and feature orientations should be avoided). Procrustes analysis [6] returns the optimal rotation and translation by decomposing the 3×3 correlation matrix between points. An early rejection of samples based on the vector length between point pairs can be employed in a similar way to our previously mentioned 2-point pose verification.

6. Experimental Evaluation

For evaluation we recorded 3 different datasets with different scene characteristics which are typical for laser scanning scenarios. CHURCH is an indoor dataset of an old church consisting of 5 scans and exhibiting many vaults. Besides peaks in the normal distribution, in this scenario we also extract symmetry planes. Note that there exists a sign ambiguity for the symmetry plane normal, thus we use both possible normal directions as salient direction. For CITY we captured 3 scans in an urban area showing a high number of structured facades (e.g. balconies). Finally CASTLE combines a construction site and a historic building [3].

For all experiments the input data format and parameters of our algorithm were kept constant. Panoramic images are and range data is represented as 6 faces of a cube-map, each of size $2k \times 2k$ and $1k \times 1k$ pixels, respectively. For salient direction estimation we subsample the depth data as explained, while the kernel bandwidth and standard deviation (Eq. 6) are set to 10° and 5°, respectively.

Repeatability of Salient Directions It is essential for successful registration that we extract at least one salient direction (up to small variation) in both viewpoints. This task becomes more difficult with less overlap between regions. For evaluation we have taken scans with known relative pose and rendered the source scene into the viewpoint of the target scene. There we compare the original depth values to those of the rendering. Areas with small difference in depth are considered as visible in both scenes, i.e. they define the area of overlap between scans. Thus, in these regions corresponding salient directions (defined as directions differing by 10° at maximum) can and should get support. We now determine repeatability scores by comparing the number of corresponding salient directions to the total number of detected salient directions. The lower left parts in Table 1 list our evaluation of repeatability scores. One can observe that re-detection rates are high.

Registration performance To demonstrate the registration performance of our approach we compare it against state-of-the-art planar RGB-D rectification [22, 4]. We also tried to match SIFT features extracted from the original images (i.e. cube face images), but registration fails in more than half of the cases. Tab. 1 lists the number of correct matches vs. tentative correspondences for both our approach and the baseline. A match is seen as correct if the corresponding points are within a threshold of 5cm for the outdoor datasets and 3cm for CHURCH (since it has smaller

[3] The datasets and additional results are available at http://www.cvg.ethz.ch/research/saldir-rgbd-registration

Figure 5: Comparison between viewpoint normalization via SDR-images (left) and planar rectification [22, 4] (right). Clearly our approach can handle arbitrary surface shape and extract features on those.

scale). As can be seen, we generate more tentative and correct matches, which enables us to register scan-pairs in cases where the other approach fails. As expected this is the case for scenes with numerous non-planar surfaces, such as the roof and apse dome in Fig. 2. Here our approach is crucial for successful registration, as planar rectification requires textured planes, which are small or non-existent (cf. Fig. 5). Note that besides exploiting features on free-form surfaces, we completely separate stable geometries and textures; e.g. salient directions can be established from a untextured white wall, while the features for matching originate from some other textured free-form surface.

In addition Fig 1 and Fig 6 illustrates the global registration results for CHURCH and CITY, respectively. Previously pair-wise estimated relative poses form a graph connecting the scans with successful registration. An initial solution for the absolute pose of each scans is obtained by construction of a minimum spanning tree (MST) in the graph and concatenating relative transformations accordingly. To improve this initial set of poses one can examine e.g. pose-graph optimization or bundle-adjustment. We execute the former but refer for details to [13], since the focus of this work is on the initial pair-wise registration. However, we want to point out that estimated relative poses are very precise as the solution obtained via a MST is very close to the solution obtained after global optimization.

7. Conclusion

In this work we have presented the novel concept of obtaining viewpoint invariance by means of an orthographic projection along detected salient directions in range data. We have proven that resulting salient direction rectified (SDR) images for corresponding salient directions in different scans are identical up to a 2D similarity transformation in the general case or even more restricted in special, but common cases. This allows to exploit texture and features not only on parametric objects like planes, cones or cylinders, but on any free-form surface in the scene. We have proposed to utilize modes in the distribution of surface normals for salient direction detection. Compared to model fitting approaches for the parametric surfaces, estimating modes via mean-shift is robust, which is reflected by the high repeatability scores we achieve. We have evaluated the algorithm on challenging scenes with wide baseline and little overlap and demonstrated superior registration performance. Future work will explore fully automatic registration of scans taken at different points in time or in different lighting, seasons or weather conditions.

Acknowledgments This project was funded by the CTI Switzerland grant #13086.1 PFES-ES 4DSites. We also like to thank Pascal Werner (ETH Zurich) and Michael Wolf (Convent St. John) for providing the CHURCH dataset.

References

[1] G. Baatz, K. Köser, D. Chen, R. Grzeszczuk, and M. Pollefeys. Leveraging 3d city models for rotation invariant place-of-interest recognition. *IJCV*, 2012. 2, 4

[2] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE TPAMI*, 1992. 1, 2

[3] Y. Cao and J. McDonald. Improved Feature Extraction and Matching in Urban Environments based on 3D Viewpoint Normalization. *Comp. Vision a. Image Underst.*, 2012. 2

[4] Y. Cao, M. Yang, and J. McDonald. Robust Alignment of Wide Baseline Terrestrial Laser Scans via 3D Viewpoint Normalization. In *Workshop on App. of Comp. Vision*. IEEE, 2011. 2, 6, 7, 8

[5] D. Comaniciu and P. Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE TPAMI*, 2002. 4

[6] D. Eggert, A. Lorusso, and R. Fisher. Estimating 3-D Rigid Body Transformations: A Comparison of Four Major Algorithms. *Machine Vision and Appl.*, 1997. 6

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conf., Manchester*, 1988. 5

[8] S. Holzer, J. Shotton, and P. Kohli. Learning to Efficiently Detect Repeatable Interest Points in Depth Data. In *Proc. ECCV*, 2012. 2

[9] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE TPAMI*, 1999. 2, 6

[10] B. King, T. Malisiewicz, C. Stewart, and R. Radke. Registration of multiple range scans as a location recognition problem: hypothesis generation, refinement and verification. In *Int. Conf. on 3-D Digital Imaging and Modeling*, 2005. 2

[11] K. Köser and R. Koch. Perspectively Invariant Normal Features. In *ICCV, Works. on 3D Repr. and Rec.*, 2007. 2

[12] K. Köser, C. Zach, and M. Pollefeys. Dense 3d reconstruction of symmetric scenes from a single image. In *Pattern Recognition*. 2011. 4

[13] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g2o: A general framework for graph optimization. In *Proc. ICRA*, 2011. 7

[14] J.-F. Lalonde, S. G. Narasimhan, and A. A. Efros. What do the sun and the sky tell us about the camera? *IJCV*, 2010. 4

[15] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 2004. 2, 5

[16] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A Comparison of Affine Region Detectors. *IJCV*, 2005. 2

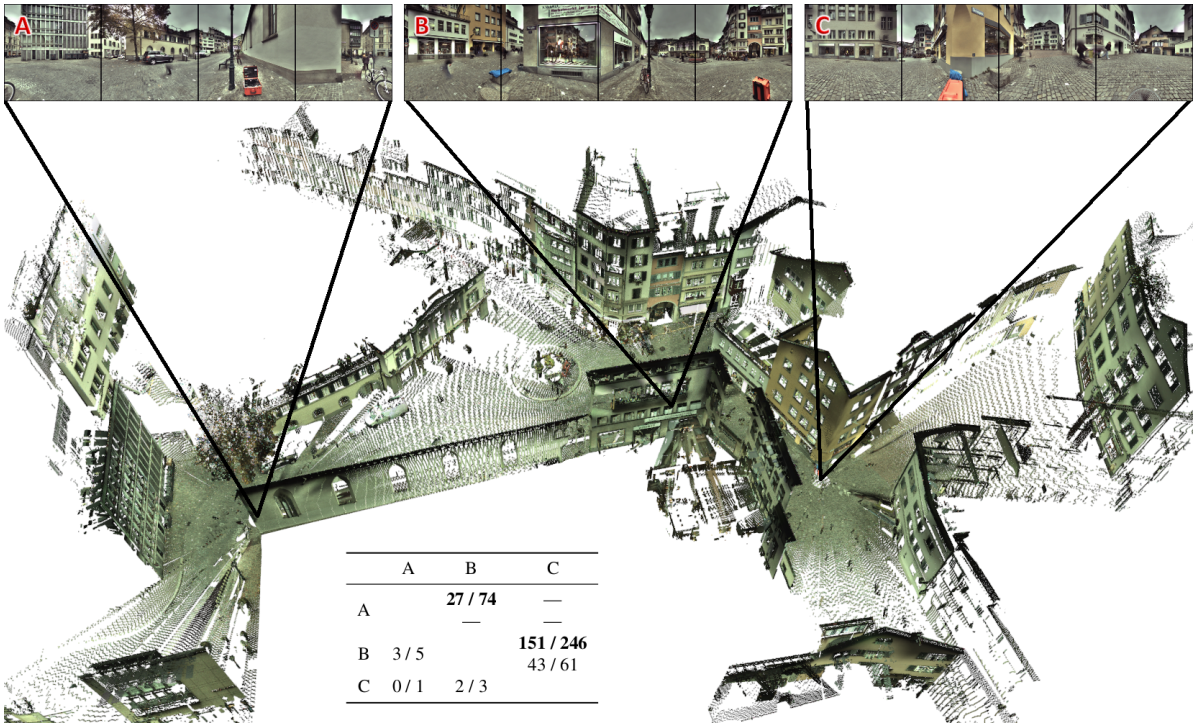
	A	B	C	D	E
A		**418 / 541** 222 / 261	**301 / 439** 127 / 160	**21 / 68** —	**15 / 39** —
B	5 / 5		**242 / 322** 131 / 161	**19 / 54** —	**53 / 95** 58 / 75
C	4 / 5	4 / 5		**159 / 225** 89 / 103	**154 / 190** 29 / 32
D	1 / 1	2 / 3	5 / 7		— —
E	1 / 2	1 / 2	2 / 2	0 / 0	

(a) CASTLE

	A	B	C	D	E
A		**335 / 419** 166 / 206	**75 / 146** —	**82 / 144** 65 / 75	**24 / 63** 16 / 23
B	6 / 7		**405 / 480** —	**349 / 435** 114 / 142	**69 / 148** 44 / 60
C	6 / 7	7 / 9		**121 / 168** —	**63 / 118** —
D	7 / 7	8 / 8	6 / 8		**123 / 166** 77 / 79
E	5 / 6	5 / 8	5 / 7	6 / 8	

(b) CHURCH

Table 1: Registration evaluation for CASTLE and CHURCH. (Upper right parts) Relation between correct and tentative matches, for our approach (in bold) and planar rectification [22, 4]. The results indicate our superior performance. (Lower left parts) Repeatability scores for salient directions, i.e. the ration of found and present salient directions in the scan overlap.



	A	B	C
A		**27 / 74** —	— —
B	3 / 5		**151 / 246** 43 / 61
C	0 / 1	2 / 3	

Figure 6: Result of our automatic scan registration for CITY. The scene was created from 3 viewpoints, visualized via their horizontal cube faces. The numbers in the table are organized in the same way as in Table 1.

[17] G. Pandey, J. McBride, S. Savarese, and R. Eustice. Automatic Targetless Extrinsic Calibration of a 3D Lidar and Camera by Maximizing Mutual Information. In *AAAI Conf. on Artificial Intelligence*, 2012. 2

[18] D. Robertson and R. Cipolla. An Image-Based System for Urban Navigation. In *Proc. BMVC*, 2004. 2

[19] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Proc. ICRA*, 2009. 2

[20] A. Telea. An Image Inpainting Technique Based on the Fast Marching Method. *J. of Graphic Tools*, 2004. 5

[21] R. Unnikrishnan and M. Hebert. Fast extrinsic calibration of a laser rangefinder to a camera. Technical Report CMU-RI-TR-05-09, 2005. 2

[22] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys. 3D Model Matching with Viewpoint-Invariant Patches (VIP). In *Proc. CVPR*, 2008. 2, 5, 6, 7, 8

[23] J. V. Wyngaerd and L. V. Gool. Automatic crude patch registration: toward automatic 3d model building. *Comput. Vis. Image Underst.*, 2002. 2

[24] S. Yamany and A. Farag. Surface Signatures: An orientation independent free-form surface representation scheme for the purpose of objects registration and matching. *IEEE TPAMI*, 2002. 2

[25] B. Zeisl, K. Köser, and M. Pollefeys. Viewpoint Invariant Matching via Developable Surfaces. In *ECCV, Workshop on Consumer Depth Cameras*, 2012. 2